IOP Concise Physics | A Morgan & Claypool Publication

Kyle Forinash

PHYSICS AND THE ENVIRONMENT

IOP | ebooks

Physics and the Environment

Physics and the Environment

Kyle Forinash III
Indiana University Southeast

Morgan & Claypool Publishers

ISBN 978-1-6817-4493-3 (ebook)
ISBN 978-1-6817-4492-6 (print)
ISBN 978-1-6817-4495-7 (mobi)

DOI 10.1088/978-1-6817-4493-3

Version: 20170501

IOP Concise Physics
ISSN 2053-2571 (online)
ISSN 2054-7307 (print)

A Morgan & Claypool publication as part of IOP Concise Physics
Published by Morgan & Claypool Publishers, 40 Oak Drive, San Rafael, CA, 94903 USA

IOP Publishing, Temple Circus, Temple Way, Bristol BS1 6HG, UK

To my wonderful wife, Miriam, whom I love dearly.

Contents

7 Energy storage 7-1

8 Transportation 8-1

9 Climate 9-1

Preface

In 2001 I went on vacation to Alaska and had the good fortune to be able to visit quite a few glaciers, on foot, in kayaks and by small plane. I recall at one point walking along a path from the parking lot towards the face of a glacier some two miles away and seeing a sign that said, simply, '1951'. I began to pay attention to the museum pictures, tour guides and signs at other glaciers and it dawned on me that, of the many glaciers that I had seen, every single one had retreated by a substantial amount in the past 60 to 100 years. When I returned home, curious as to exactly what I had seen in Alaska, I began to investigate the scientific literature on climate change and environmental issues much more intently. I had already taught a course in environmental physics in 1999, focused mainly on energy, but I do not think the magnitude or significance of climate change had really sunk in until I saw those retreating glaciers. This gave me a much broader picture of the interconnectedness of the environment and our global dependence on fossil fuels. My technical understanding has continued to deepen over the intervening years and I hope this book conveys the significance of some of those connections.

When I was born there were about 2.5 billion people in the world. We are now at 7.4 billion. Furthermore, three quarters of the world has a much higher standard of living than they did 65 years ago. This near tripling of the population and the associated increase in per capita resource use is the basis of many environmental issues facing the world today. Technology has given us vaccines, antibiotics and a steep increase in agricultural productivity, which is part of the explanation for the sharp increase in population. However this same technology has led to increased pollution, contaminated water, destruction of natural biological systems, antibiotic resistant bacteria, a dependence on mono-crop food sources and many other undesirable environmental effects. Chapter 1 of this book is an introduction to population dynamics and the associated problems for the environment.

From my readings since 2001 I have learned that the landscape encompassing environmental issues changes quite rapidly. Few researchers 30 years ago would have predicted the current boom in natural gas due to hydraulic fracturing, or that electricity from wind energy in many parts of the US and elsewhere would become cheaper than coal. These facts were not completely evident just ten years ago when I wrote my first book on environmental physics[1]. However the physical principles needed to understand these transformations have not changed and this idea—that known physical laws can be applied to novel physical problems in the environment—is a core theme of this book. The first and second laws of thermodynamics, introduced in chapters 2 and 3, still limit the efficiencies of energy conversions to less than 100%. Heat engines will always be limited by the Carnot cycle efficiency. However, sometimes clever technology can enhance overall performance. Capturing the forward kinetic energy of a slowing vehicle to be used later, examined in chapter 8,

---

[1] Some of the material in this book is adapted (with permission) from my earlier book, *Fundamentals of Environmental Physics*, Island Press, 2010, Copyright 2010, Kyle Forinash.

is an example where the same fundamental concepts can be used to evaluate and understand an old problem and extended to study a new solution, and perhaps provide tools to improve on it.

Technologies using fossil fuel resources, the topic of chapter 4, are, for the most part, mature and well developed. We are unlikely to see huge gains in the efficiency of coal or natural gas power plants. Hydraulic fracturing and the retrieval of petroleum from tar sands are examples of the application of new technology to the retrieval of existing energy resources, but at the cost of a much smaller energy return on energy invested in recovering those resources. The world still has dozens of years of fossil fuels left, perhaps as many as 100 years, but eventually these sources will have to be replaced. Their role in climate change ought to be an argument for their replacement sooner rather than later.

Renewable resources, discussed in chapter 6, mostly involve new technologies which are still under intense development. Cheap, commercially available photovoltaic solar cells have efficiencies of less than 20%, but laboratory efficiencies are reaching 50% and these new technologies will reach the market in the near future. But in order to make a smooth transition between non-renewable and renewable energy it is important to become aware of many physical constraints. How many new windmills or photovoltaic cells and with what efficiency and output are needed to entirely replace global electricity from coal or natural gas? The answer depends on the available technology but it is not a trivial number. Renewables, including hydroelectricity, sill count for less than 20% of the global primary energy supply, so if we want to phase out fossil fuels in the near term we need to start building more renewable systems soon, and lots of them.

Do we have an efficient and large enough means to store renewable energy so that it is available when renewable resources are not (chapter 7)? It should be clear given the low energy density of all renewable resources that energy storage is a key factor in consuming significant amounts of renewable energy. This is a core question which remains unanswered but must be solved using the tools of science and engineering, not opinion or politics.

Another core element of this book is to introduce some of the technical issues surrounding nuclear energy (chapter 5). Some parts of the world (China) are busy building dozens of nuclear power plants while others (Japan and Germany) are seeking to move away from nuclear energy as fast as possible. Other places remain undecided (the US). Having a basic understanding of how nuclear power plants function is a prerequisite for any citizen to have an intelligent, well-formed opinion either for or against nuclear power.

The basic features of why we have a climate (chapter 9) are easily understood from a physics perspective, although there are many subtleties in the details. It is clear from very simple calculations that changing the amount of carbon dioxide in the atmosphere (which humans are doing quite rapidly on a geological time scale) should cause changes in surface temperatures on the Earth or any planet or satellite with an atmosphere. Paleo-climate and planetary data show this to be the case and recent data also show a direct link between carbon dioxide and global temperatures, confirming well-understood physical theory. Unfortunately temperature is not the only effect of human caused climate change. We are just now becoming aware of many other potential problems having to do with additional carbon dioxide in the air and oceans and many of these changes may have grave consequences.

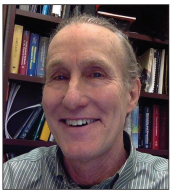
A final consideration for evaluating a potential solution to any environmental issue is a cost benefit analysis. To do this properly the physical risks must first be assessed before any economic issues are introduced. In many cases a probabilistic estimate of the physical risk is relatively straightforward and possesses a high degree of accuracy. No environmental solution, no matter how clever from a technical aspect, will succeed if there are cheaper alternatives, even if those alternatives have undesirable features associated with them. The best way to internalize the external costs of technology is often not obvious and these decisions suffer from political and social biases but a first step is to accurately evaluate the cost of those externalities and the associated risk involved. A brief overview of these issues is given in chapter 10.

Acknowledgment

I would like to thank the many students in all of my classes over the past 40+ years who have helped me think about these issues by asking interesting and thoughtful questions and then by questioning my answers. I would like to also thank Wes Halfacre for suggestions on the climate chapter. Special thanks go to Mike Moody, my lifelong fellow traveler and colleague in physics and philosophy until his death last year, and to Mike Aziz, both of whom have helped me see deeper into the physics of environmental issues over the years.

About the author

Kyle Forinash



Dr Kyle Forinash is Professor of Physics at Indiana University Southeast in New Albany, Indiana. He has published technical papers on non-linear dynamics, in addition to pedagogical papers about the use of computers and cell phones for data collection in student laboratories. He is the author of two books, one in mathematical physics and the other in environmental physics, and two electronic interactive ePubs on iTunes, one on waves and the other on the physics of sound. His current interests include applications of physics to environmental issues and open source computer simulations for physics education. A partial resume is available at http://pages.iu.edu/~kforinas/.

Chapter 1

Population and its effects

We can achieve no lasting solution to any of the major environmental problems facing modern society without a thorough understanding of population demographics. The population of the world and how it lives has transformed radically in the past 200 years, and these changes have had an enormous effect on the global environment. This chapter relates a few basic statistical facts about the global population to their impacts on natural resources and the global environment. Also included is a brief discussion of how the amount of contact with a chemical substance is determined in various situations and how to go about investigating whether there is a statistically significant outcome related to that exposure.

1.1 Population

For the first time in human history a single generation has seen the global population double in their own lifetime. There were about one billion people in the world in 1800 and it took 130 years for that number to double to two billion. It only took 45 years, starting in 1930, to reach four billion in 1975. The global population is now 7.4 billion, having added one billion people in the past 12 years (UN 2015). The exponential growth in the world's population is a major driver of today's environmental problems, from use of resources such as energy, food and water, to problems associated with pollution, waste disposal and climate change (treated in chapter 9).

Where and how people live are also changing and this compounds the problems caused by the steep increase in population[1]. In 1900 about 13% of the world's population lived in urban areas. By 2008 the number of urban dwellers passed the number of rural residents and by 2050 the percentage of the global population living

---

[1] The IPAT equation was one of the first attempts to quantify the impact of population and consumption on the environment. $I = PAT$ where I is impact, P is population, A is affluence and T is technology (Commoner 1972). There are other, similar attempts to quantify the connection between population, technology and environmental impact.

doi:10.1088/978-1-6817-4493-3ch1

in cities is expected to reach 60% (Weeks 2016). There are many implications of this demographic shift. On one hand, this change marks an increase in general wellbeing; urban populations commonly have better access to education, health services and economic opportunities, especially for women. On the other hand, a rapidly urbanizing population faces a lack of infrastructure to accommodate these new city dwellers which results in poor sanitation, lack of basic services such as electricity and water, and unsuitable living conditions. Unlike rural populations which generally grow their own food, urban populations depend on large scale monocrop farming of hybrid varietals using modern farm equipment and fertilizers to feed them. The large scale change from rural to urban was made possible in part by repurposing of nitrogen after the Second World War from making explosives to making fertilizer, resulting in a significant transformation in the way the world is fed.

Fertility rates are declining worldwide, in large part due to better access to education and economic opportunities for women. The *general fertility rate* is defined as the number of live births per 1000 women of childbearing age per year and the *total fertility rate* is the average number of children born per woman. Depending on the reporting source, childbearing ages may be assumed to be either 15–44 or 15–49. Due to infant mortality, the *replacement rate* (total fertility rate needed for the population to remain static in number) is 2.1. Total fertility rates in the developed world have dropped from three children per woman in 1950 to less than two today. The population in the US and Europe is increasing due to immigration, not new births, and populations in Russia and Japan are declining. Total fertility rates in most of the developing world, except for the poorest countries, have fallen from six children per woman in 1950 to less than three today. Total fertility rates in some parts of Africa are still as high as five children per woman, but are projected to drop to under three children per woman by 2050. Because the effect of a decline in fertility rates does not become evident until the children who are born reach child bearing age, the global population is not projected level off until after 2100 at around 11 billion people (UN 2015).

As a result of modern medicine people are also living longer, as can be seen by the death rates shown in figure 1.1. The effect of modern medicine on children under one year of age is particularly evident. Part of the increase in the global population is the result of increased longevity and a healthier population, both of which have added effects on society. The median age of the world's population in 1950 was 24 but today about half the global population is over the age of 30. By 2050 it is projected that the global median age will be 37 with a median age of 50 in developed countries. Older persons in general are not as productive as younger people and also need more health care resources, adding to environmental problems.

There are now twice as many mouths to feed, twice as many beds needed, twice as much housing, twice as many jobs needed, twice as much energy needed to stay warm or cook, and twice as any articles of clothing needed due to a doubling of the population in the last 60 years. If we still lived at the same level of consumption as 1960 we would need twice as many resources but because of shifting demographics and a higher standard of living globally, we are using many more resources than twice the amount used in 1960. Currently about 12% of the ice free land surface of

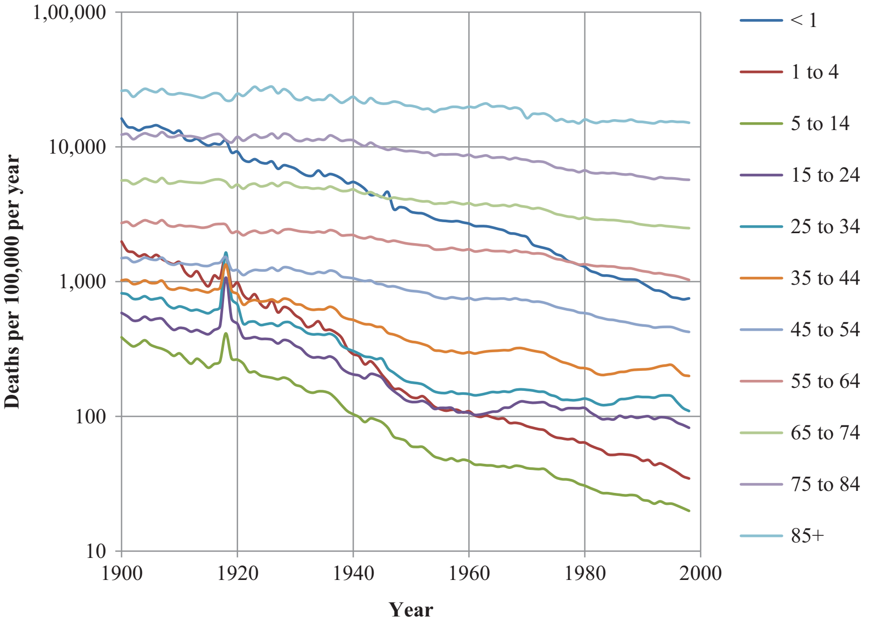


Figure 1.1. Historical death rates in the US by age (CDC 2016). This chart will be discussed in more detail in chapter 10.

the earth is farmed (Smil 2015). Estimates of the human appropriation of the global net primary productivity (NPP—the total amount of plant mass photosynthesized in a period of time, usually estimated as annual growth, minus respiration losses) range from 3% up to 55% depending on the methods and assumptions used (Smil 2015). The increase in population coupled with changes in the way most of the world lives has had an outsized impact on the natural environment[2].

1.2 Water and food

Prior to 1900 most water was used for farm irrigation and this is largely still the case in developing countries. The developed world uses a much larger fraction of extracted water for industry and domestic use and this has increased at a much faster rate than the population. Personal water use has increased by about 25% over the past 100 years but total water extraction has increased by a factor of eight (McNeill 2000). Per capita water for personal use in the US is highest globally at 217 m^3 per person per year compared to 38 m^3 for the average citizen of India and 26 m^3 for the average Chinese citizen (Mekonnen and Hoekstra 2011), but the total water withdrawal (the *water footprint*) and associated stress on renewable water resources of China and India is higher than the US because of the larger number of people living there (National Water Footprint 2016).

---

[2] Approximately one billion people in the world still live at about the same level of development as 200 or more years ago. However, the other six billion are much better off than 200 years ago (Collier 2007, Norberg 2016).

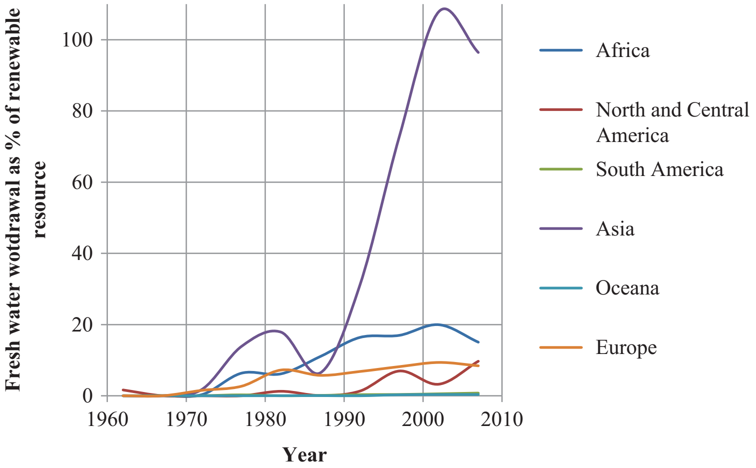


Figure 1.2. Freshwater withdrawal as a percentage of renewable water resources (FAO 2016). Asia includes the Middle East which imports water as virtual water in the form of agricultural and manufactured products.

The term *green water* is often applied to the water used in agriculture that falls as rain and from rain runoff. More than half of the world's agriculture production depends on green water sources. The term *blue water* is used for water extracted from lakes or aquifers which is used for a variety of purposes besides agriculture, with the result that there is competition for its use[3]. *Grey water* is water that has been used and so is polluted but could potentially be or is treated and reused. Many countries are currently using all or nearly all of their renewable water resources (green, blue and grey) and some countries import water. For example Egypt is using 92% of their renewable resources, Iran 48%, Pakistan 73% and Saudi Arabia 643%; in other words Saudi Arabia must import more than six times the water it has available (FAO 2016).

Imported and exported water is often in the form of *virtual water*, which is water used for growing crops or required for the manufacture of an item. For example about 140 liters of water is required to produce a cup of coffee, 3000 liters for a kilogram of rice and 22 000 liters for a kilogram of beef. Roughly 49% of the water withdrawn in the US goes to cooling thermal electric generation plants (coal, gas and nuclear); transmitting electricity across a border in effect transfers water consumption from where the electricity is generated to where the electricity is used. Saudi Arabia imports virtual water in the form of food and manufactured products which were created using water in other parts of the world; this contributes to the consumption of 643% of its renewable resource. Virtual water imports are important in the Middle East, Pakistan and North Africa as reflected in the data shown in figure 1.2, increasing the percentage of renewable resource used.

If people lived the same way they did 60 years ago we would expect a doubling of the consumption of livestock and marine products between 1960 and now to match the doubling of the population that occurred in this time frame. However, more

---

[3] For example, recreation and industrial use.

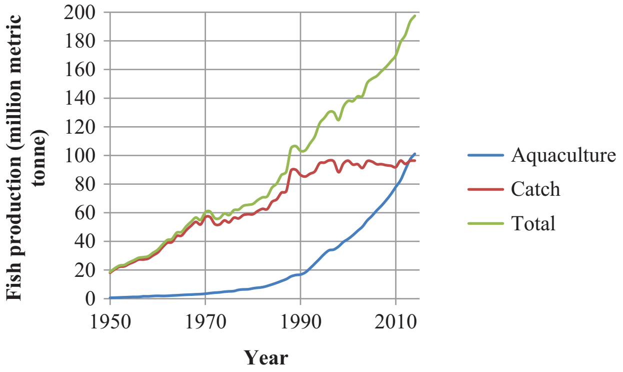


Figure 1.3. Global marine production (FAO 2016). Includes fish, shellfish and other edible ocean marine life.

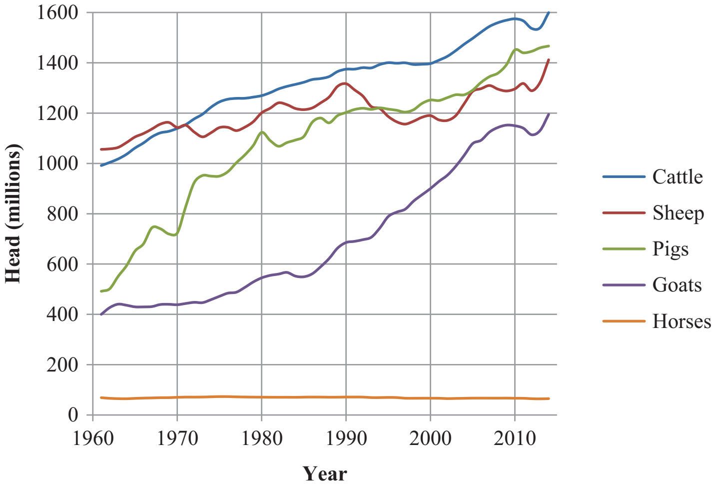


Figure 1.4. Global livestock production (FAO 2016).

people today live in cities, are healthier, live longer and eat better than they did 60 years ago. Global marine production, which includes fish, shellfish, crustaceans and other marine life has increased by a factor of eight, mostly due to aquaculture, as seen in figure 1.3.

The standard of living in China, India and some parts of Africa has increased greatly in the last 60 years resulting in a threefold increase in goat and pig production, as seen in figure 1.4. Synthetic clothing has likely had an effect on sheep production which did not double in the last 60 years. Cattle production increased by 150% during this period although cattle are a more common food supply in the developed world where population increases have been less than in the developing world. We do not eat horses and no longer use them for transportation so the global horse population has remained constant. The mass of all wild land animals existing today is outweighed by the mass of living humans by a factor of 10 and domesticated animals by a factor of 24 (Smil 2015).

Urbanization and economic development also bring broader use of modern appliances. The number of refrigerators and Internet use have saturated in the US market, remaining approximately constant since the early 2000s; everyone in the US now has a refrigerator and access to the Internet. However cell phone ownership just about doubled during this period, reflecting an expanding new market. The developing world is still catching up however. During this same time frame, per person Internet use, number of refrigerators and number of cell phones all tripled in China (UN 2013). If the developing world, which constitutes more than 70% of the world's population, achieves the same standard of living as the developed world we can expect the increase in resource use to continue to far outpace population growth over the next few generations.

1.3 Pollution

It would seem logical that increasing populations will result in more pollution and waste and this has been the case for many types of pollution, but there are a few recent exceptions. Today outdoor air pollution is estimated to result in 3.3 million premature deaths per year (Lelieveld *et al* 2015)[4]. Before regulations were put in place in the 1980s in the US, lead was used in paint as an anti-mold and mildew additive and in gasoline to make engines run smoother. This resulted in a large amount of lead entering the environment and many urban areas near busy traffic routes still have high levels of lead in the soil. Airborne metals travel globally before being washed into the soil and in fact 70% of the lead found in surface water and soil comes from air emissions. Even in Antarctica, where most air circulates in a circumpolar configuration and does not mix well with other circulation patterns, ice core data show a factor of four increase in lead contamination since 1900. As shown in figure 1.5, however, heavy metal emissions around the globe are now decreasing as the result of better regulations, especially in the developed world.

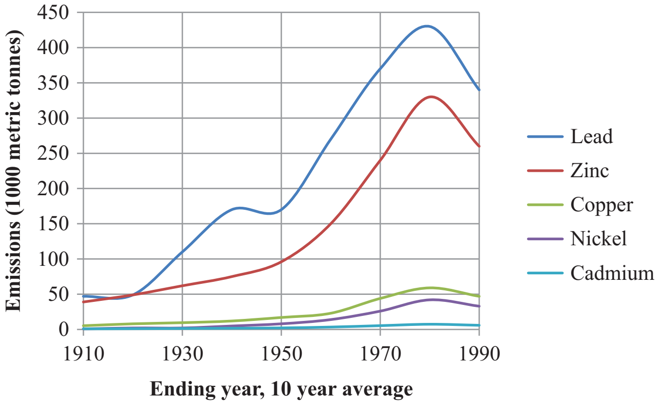


Figure 1.5. Global metal emissions of five heavy metals, not including natural sources (Nriagu 1996).

---

[4] Indoor air pollution from cooking over an open fire doubles this number.

The average per capita waste generated by a person in the US has increased to more than two kilograms of waste per day when industrial waste for manufacturing consumer goods is included. This increasing per capita waste stream has led to a tripling of municipal waste in the US since 1960, a higher rate than the doubling of the population. Recycling in the US rose dramatically between 1960 and 2000, for the most part because of economic incentives, but has leveled off since. Roughly a third of the collected municipal waste is recycled in the US, 60% in Europe. Automobile batteries are currently recycled at a rate of nearly 99% in the US because of safety regulations and the fact that all of the material can be re-used in making new batteries. Steel cans and yard waste are recycled at a rate of 60%, aluminum cans at 45%, paper at 51%, and 30% for plastic and tires. The US municipal waste stream is characterized in figure 1.6. There are significant energy savings in recycling most metals (aluminum in particular because the pure metal's separation from bauxite ore is energy intensive) whereas paper and plastic recycling use nearly the same energy as it takes to make the original product. Plastic recycling is often problematic because different types of plastic melt at different temperatures and have different thermal properties. A single bottle of the wrong type of plastic can ruin an entire batch of a given type of plastic being melted down for a recycled resource stream.

Electronic equipment, such as computers, cell phones, microwaves and television sets, contains valuable metals which would seem to make them attractive for recycling, however the procedure involved in separating these metals from their components is currently done by hand and so is labor intensive. This makes recycling electronics economically unattractive in the developed world where wages tend to be high. As a result, a number of West African countries, where labor is cheap, have become dumping grounds for electronic waste from the developed world to be salvaged. Agbogbloshie, Ghana, is reported to be one of the most polluted places on Earth because it has the second largest electronic waste dump in the world. Electronic cast offs from developing countries are broken apart by hand and burned in open fires to extract the precious metals inside, contaminating the surrounding

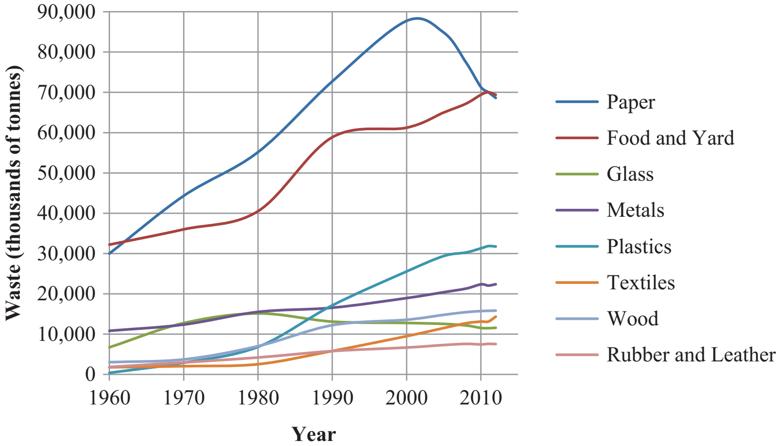


Figure 1.6. US municipal waste stream (EPA 2014a).

environment with toxic heavy metals. This same type of process, performed by unprotected low paid workers who are not aware of the danger of exposure to these contaminants, occurs in many other developing countries as well (Biello 2013).

Alternatives to recycling include landfills and incineration but neither of these options is without problems. Incineration can be used to generate energy but the plants are costly to build. Much of the available land for landfills is now in use in the developed world making it difficult to find enough room for the enormous amounts of garbage being disposed of today. Both landfills and incineration lead to heavy metals, dioxin and other contaminants being released into the environment.

An increasing amount of waste is being washed into rivers and oceans, significant amounts of which are made of plastic which does not degrade rapidly. Ocean currents sometimes push debris all the way across an ocean, as happened after the Japan 2011 tsunami where some floating material, including living plants, crossed the Pacific Ocean within a few months. Ocean circulation patterns can also form gyres or convergence zones where trash tends to accumulate. These so called ‘garbage patches’ form in all the oceans and the plastic there gradually breaks apart into tiny pieces that can be found in the upper few meters of the ocean (van Sebille *et al* 2015). These colorful pieces sometimes harm bird and sea life which mistake them for food. The garbage patch name is somewhat misleading as the increase in plastic content by volume compared to pure water is relatively small but all oceans today contain significant amounts of plastic (Litterbase 2017).

Coal was the first widely used fossil fuel and increase in the use of coal quickly led to serious local pollution problems. In the London Fog episode of 1952 a temperature inversion trapped fog and smoke in London leading to a sudden increase in respiratory problems, including death. It is estimated that approximately 12 000 people died as a result of the sudden increase in pollutants in this single week-long event. This unintended experiment was among the first to bring wide public attention to the fact that air pollution is a significant health problem, not just a nuisance (Nielson 2002). The burning of fossil fuels, coal in particular, constitutes one of the largest sources of air pollution today.

Initial attempts to ameliorate local pollution from fossil fuel burning involved building taller smoke stacks so that the pollution was spread over a larger area. The taller stacks had the effect of changing the pollution from a local problem to a regional problem, particularly for sulfur emissions which cause *acid rain*. The burning of coal to generate electricity in the US is responsible for about 65% of the sulfur dioxide (SO_2) emissions and 30% of the various harmful nitrogen compounds (NO_x where x is a whole number) found in the atmosphere. The exact amount of sulfur emitted per tonne of coal burned depends on the type of coal used, but generally some sulfur is always present. Once oxidized into sulfur dioxide during the burning process the sulfur compounds undergo the following reactions in the atmosphere:

$$\begin{aligned} 2SO_2 + O_2 &\rightarrow 2SO_3 \\ SO_3 + H_2O &\rightarrow H_2SO_4. \end{aligned} \tag{1.1}$$

Sulfuric acid, H_2SO_4, is in liquid droplet form and disassociates into the ions $2H^+$ and SO_4^{-2} which can react with plants and building material as acid rain. The same reactions can occur for sulfate deposited directly onto the ground as particulates, a process called *dry deposition*.

Acidity is measured as pH which is defined as the log of the hydrogen ion concentration,

$$\mathrm{pH} = -\log[\mathrm{H}^+]. \tag{1.2}$$

Pure water has a H^+ concentration of 10^{-7} moles per liter and so has a pH of seven. Liquids with a pH lower than seven are acids, those with higher pH are bases. Some lakes in the northeastern part of the US have been measured to have a pH as low as three (about that of vinegar) as the result of acid rain, killing all or most of the life in them. The water appears very clear and clean but this is because no microorganisms, which give large bodies of water most of their color, can live there.

The US the Environmental Protection Agency (EPA) put a *cap and trade* program in place under a 1990 modification of the 1970 Clean Air Act whereby emitters could buy and sell permits to emit fixed quantities of sulfur, thus creating a market incentive to reduce emissions. This program has been very successful (a 69% reductions in sulfur dioxide between 2005 and 2013) and demonstrates how the right legislation can reduce pollution substantially (EPA 2016).

We often think of pollution problems as limited to geographic regions but contamination of the air is a global phenomenon. Upper atmospheric air currents move particulates including dust, bacteria, sulfur compounds and other pollutants in a global circulation pattern. This has led to large areas of the globe being covered by smog-like clouds containing industrial pollutants which can be measured from space, for example the reoccurring Asian Brown cloud over India and the Indian Ocean (Ramanathan *et al* 2001). Figure 1.7 shows a NASA projection of dust moving from the Sahara towards Central and North America. The particulates include bacteria and other microorganisms which have been collected in measurable quantities in North America.

Dust is not the only thing that blows around the globe in significant quantities. Figure 1.8 shows a forecast from satellite data and computer models of carbon monoxide arriving in North America caused by biomass burning in Eastern Europe and Asia. Similar projections and data exist for sulfur dioxide, carbon, ozone and particulates, including microorganisms. We cannot solve environmental contamination problems without considering them from a global perspective.

Some 80% of North Americans and 60% of Europeans live in locations where the Milky Way cannot be seen at night due to light pollution (Falchil *et al* 2016). The effects of too much artificial light on insects and animals has been established for many species (Rich and Longcore 2005). Examples include sea turtles hatching at night near the ocean which need a dark sky to orient properly and migratory birds being confused by tall building with lights. It is now known that humans have non-imaging sensors in the eye implicated in setting a circadian cycle of sleep and awake times. It is believed that artificial blue light, prevalent in screens of electronic devices, can affect this photo-sensor and the circadian rhythm it helps to control.

Similar to light pollution, our modern society exposes us to sound intensities much higher than those with which we evolved. Sound intensity level (SIL) is a

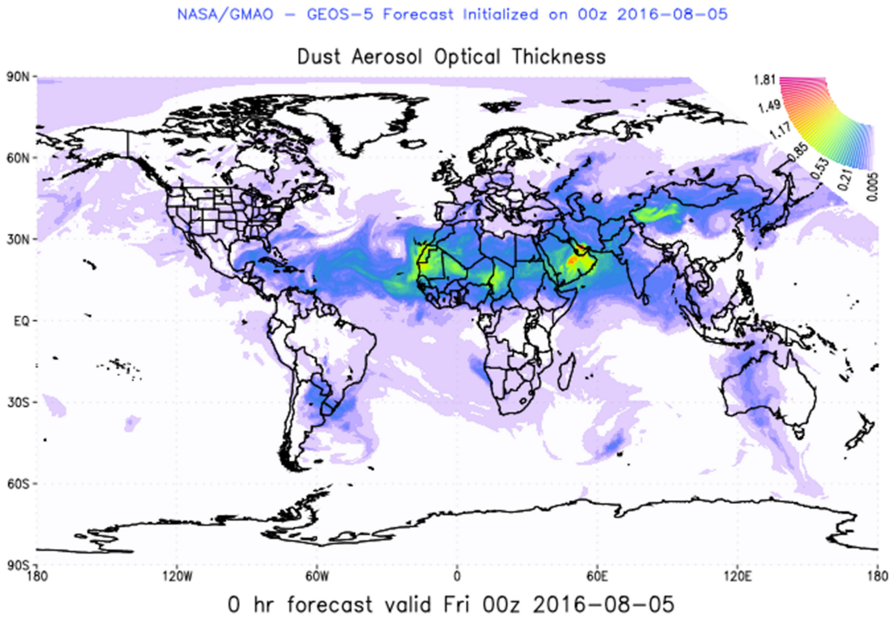


Figure 1.7. Global particulate forecast based on satellite data and computer models. Dust from the Sahara is moving westward and being deposited in Central and North America. Dust from Saudi Arabia is moving south, and dust from the western Gobi desert is moving southwest into Pakistan. Animated maps are available from NASA (2016). (NASA open source image.)

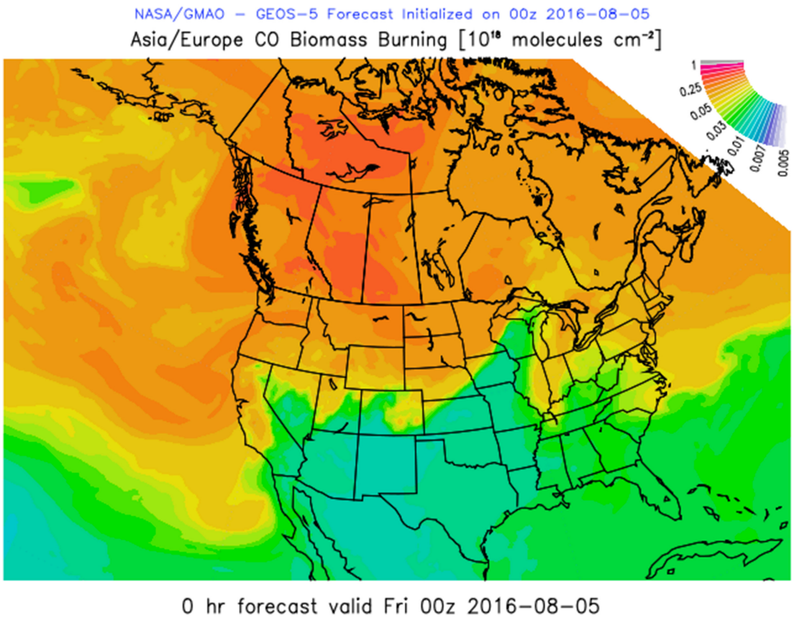


Figure 1.8. Forecast of carbon monoxide emissions caused by biomass burning in Asia and Eastern Europe, arriving in North America from the west. Animated maps are available from NASA (2016). (NASA open source image.)

logarithmic scale in decibels (dB) measuring the energy arriving per second in a square meter area. The definition is $\mathrm{SIL} = 10 \log \frac{I}{I_0}$, where I is energy intensity of the sound in W m^{-2} and I_0 is a reference energy intensity of 10^{-12} W m^{-2}. I_0 is considered the threshold of human hearing and represents the energy of a pressure wave with rms[5] pressure variations of 0.000 02 Pa. Normal conversation occurs at about 60 dB but we are exposed daily to TVs (70 dB), rock bands (110 dB), car horns (120 dB, which is considered the threshold of pain) and jet engines (140 dB) which denotes an rms pressure variation at the eardrum of 200 Pa. Both wildlife and humans experience many detrimental effects from over exposure to loud sounds, such as hearing loss, stress and cardiovascular deterioration over time. There has been a rise in hearing loss among adolescents in the past few decades, most likely as the result of over exposure to personal music devices (Shargorodsky *et al* 2010). It is not clear how to persuade ear-bud users to change habits which would avoid this damage.

1.3.1 Pollution scaling

General consensus agrees that larger cities tend to be more polluted in general than smaller ones although one might rationally argue that for pollution sources which are distributed uniformly per unit area the amount of pollution would *not* depend on the size of the city. Scaling laws come into play leading to a result more familiar to our experience even if the source is proportional to area.

We start with two rectangular shaped cities, one with a length on each side of $L_{\mathrm{small}} = 4$ km and the other with a side $L_{\mathrm{big}} = 40$ km. Assume both cities have the same distribution of pollution sources per area, S. The ratio of their lengths is 40/4 = 10 or $L_{\mathrm{big}} = 10L_{\mathrm{small}}$, however their areas scale as $1600\ \mathrm{km}^2/16\ \mathrm{km}^2 = 10^2$ or $A_{\mathrm{big}} = 10^2 A_{\mathrm{small}}$. The rate of pollution production, $R_{\mathrm{production}}$, for each city will be the source per area, S, times the area, or $R_{\mathrm{production}} = SA$, so we have

$$R_{\mathrm{production(big)}} = 10^2 R_{\mathrm{production(small)}}. \tag{1.3}$$

This would indicate that the cities have the same pollution density because the big city is 100 times the area of the smaller one, however, other factors come into play to change this conclusion.

Industrial pollutants and contamination from electricity generation tend to be found in warmer air which will cool as it rises until it is the same temperature as the surrounding atmosphere, after which it will not rise any further. This height is called the inversion height, h, and depends on the geographical terrain, temperature and weather patterns, but for the most part does not depend on the size of the city.

Each city would have the same amount of pollution if the rate of removal was the same but pollution removal scales as the length of the side of the city and depends on air speed, concentration and inversion height in the following way. Wind at velocity v will blow pollution away from a city through a vertical cross sectional area of $L \times h$ where L is the length of one side of the city and h is the inversion height.

---

[5] Root mean square (rms) is the square root of the mean amplitude squared of an oscillating wave.

The flux of pollution removal, f, in kg $s^{-1}m^{-2}$ at one edge of a city is $f = vc$, where c is the concentration of the pollutant in kg m^{-3} so the rate of removal is $R_{\text{removal}} = vcLh$ in units of kg s^{-1}.

On reaching equilibrium all the pollution in each city eventually gets blown away so that the rate of production equals the rate of removal for each city, we have

$$R_{\text{production(big)}} = R_{\text{removal(big)}} = vc_{\text{big}}L_{\text{big}}h \tag{1.4}$$

and

$$R_{\text{production(small)}} = R_{\text{removal(small)}} = vc_{\text{small}}L_{\text{small}}h.$$

Taking the ratio of these two equations and using the area scaling from equation (1.3) we have

$$\frac{R_{\text{removal(small)}}}{R_{\text{removal(big)}}} = 10^2 = \frac{vc_{\text{small}}L_{\text{small}}h}{vc_{\text{big}}L_{\text{big}}h}. \tag{1.5}$$

Using the length ratio $L_{\text{big}} = 10L_{\text{small}}$ and canceling like terms we have $c_{\text{big}} = 10c_{\text{small}}$; in other words the pollution concentration in the big city is ten times the amount in the small city, even if each city has the same distribution, S, per area of polluting sources (adopted from Hafemeister (2007)).

1.3.2 Ozone

Surface level or troposheric ozone (O_3) is a strong oxidant that can cause respiratory problems in animals, including humans, and damage plant growth. Automobile exhaust and gas tank vapors are the largest source of CO, non-methane volatile organic carbons (NMVOCs) and harmful nitrogen compounds (NO_x where x is a whole number representing different chemical structures) in the atmosphere and photochemical oxidation of these compounds is the major source of surface level ozone. Various strategies have been implemented in most developed countries to reduce the amount of harmful surface level ozone, including gas pump collection devices and local government recommendations to fill automobile tanks at night so the petrol vapors have time to dissipate before being acted on by sunlight.

Stratospheric ozone is created by ionization of atmospheric molecules starting in the lower stratosphere below 50 km and protects the Earth's surface from harmful ultraviolet (UV) radiation. This ozone forms in a process which can be summarized by the following reactions:

$$hf_C + O_2 \rightarrow O + O \rightarrow O_2 \tag{1.6}$$

and

$$O + O_2 + M \rightarrow O_3 + M,$$

where $hf_C = hc/\lambda$ represents the energy of a UVC photon (wavelengths between 220 and 290 nm) with frequency f_C. Here c in the expression hc/λ is the speed of light and h is Planck's constant, 6.6×10^{-34} Js, and λ is wavelength of the photon. The symbol M represents any other available molecule (for example molecular nitrogen or oxygen) and is required for momentum conservation when these molecules collide.

When UVB radiation (with wavelengths between 280 and 320 nm) encounters an ozone molecule it causes the ozone to disassociate into individual oxygen molecules. The resulting chemical reactions are

$$\begin{gathered} hf_B + O_3 \rightarrow O_2 + O, \\ 2O_3 \rightarrow 3O_2 \end{gathered} \tag{1.7}$$

and

$$O + O_3 \rightarrow 2O_2,$$

where hf_B is a UVB photon.

The two above reactions partially block medium (UVB) and completely high energy (UVC) UV radiation from reaching the Earth's surface. This high energy UV light kills phytoplankton in the ocean, which is an important part of the marine food chain, and can kill or cause damage to living organisms by causing cataracts and skin cancer. Low energy UV radiation (UVA with wavelengths between 320 and 400 nm) passes through the atmosphere and is useful to living organisms, activating the biological synthesis of vitamin D in humans. Life on Earth did not colonize land surfaces until 420 million years ago when sufficient ozone was created in the upper atmosphere from oxygen generated by photosyntheses in the ocean to block the damaging effects of UVB and UVC.

Until the 1930s the reaction shown in the equations, plus a related reaction involving nitrogen oxide (N_2O) and nitric oxide (NO), led to a stable layer of ozone of varying thickness in a region between 15 and 40 km above the Earth's surface with a mass of around 5×10^9 tonnes. The early 1930s saw the introduction of the use of compounds of carbon, fluorine, bromine and chlorine, called chlorofluorocarbons (CFCs), as refrigerants, fire extinguisher propellants, solvents, aerosol propellants, dry cleaning fluids and foam blowing agents. These compounds are remarkably stable, chemically, and remain in the atmosphere for dozens of years.

In the 1960s it was discovered that these long lived CFCs were reaching the stratosphere and participating in the following reactions:

$$\begin{gathered} hf_B + CF_2Cl_2 \rightarrow CF_2Cl + Cl, \\ Cl + O_3 \rightarrow ClO + O_2, \end{gathered} \tag{1.8}$$

and

$$ClO + O \rightarrow Cl + O_2,$$

where f_B is a UVB photon and CF_2Cl_2 is a typical CFC. Because the chlorine molecule remains at the end of the reaction it continues to act as a catalyst for the destruction of more and more ozone. One CFC molecule is estimated to cause millions of ozone molecules to turn into molecular oxygen, O_2. This process was first measured to have a significant effect over Antarctica where a hole in the ozone layer was discovered and it was thought that, should this process continue, the protective layer of ozone in the upper atmosphere would eventually disappear entirely.[6]

---

[6] An ozone hole does not occur over the Arctic because the Arctic is an ocean surrounded by land which promotes the mixing of the upper layers of the atmosphere. The ocean surrounding Antarctica inhibits atmospheric mixing, resulting in a cooler upper atmosphere where these reactions are more likely to occur.

One encouraging success story of international cooperation on environmental issues was the multinational signing of the Montreal Protocol in 1987 which restricted the global use of CFCs and arranged for the phasing out of these compounds. The ozone hole continued to increase until very recently because of CFCs already in the air and their long lifetime. The hole fluctuates in size over the year but the maximum size over the past few years has been stable and current measurements show it may be starting to close, but this will not be definitively proven for several more years.

1.4 The effects of pollution

Epidemiology is the study of patterns, causes and effects of disease in a given population. In many cases this involves a calculation of the risk associated with exposure to different doses of a given contaminant based on a combination of animal studies and statistically analyzed historical data. Matching the effects of a given exposure to a particular disease is often a complex task. For example, the effects of smoking were not initially obvious because there is about a 20 year lag, on average, between when a person starts to smoke and when they show signs of lung cancer or cardiovascular disease. Although the exact mechanism of how smoking causes cancer is still not completely understood, the statistical data are strong enough to conclude that smoking causes cancer, the result of many epidemiological studies. The complexity of the effects of different dosages is one reason the much lower dose of carcinogens in second hand smoke has only recently been decisively identified as a cancer agent. Often animal studies are used to validate and extend data first revealed by epidemiological studies. Ideally there should be information from both animal studies and statistical data to make a decision about limiting exposure to a contaminant. In fact epidemiological data by themselves can never establish cause, only a strong cause and effect relation, because they do not deal directly with the underlying biochemistry.

Table 1.1 shows the life-cycle emission of six key atmospheric pollutants from several alternative electricity production sources. Carbon dioxide and methane are greenhouse gases, sulfur dioxide is a source of acid rain, and volatile organic compounds and particulates[7] cause asthma and respiratory problems in humans and animals. Mercury emissions, not shown in the table, are highest for coal fired electrical plants. Knowledge of the adverse health effects of these contaminates comes from epidemiological studies.

As mentioned previously, one of the earliest connections made between health and air pollution occurred in London in 1952. In the first week of December of that year the level of smoke and soot rose from about 500 $\mu g\ m^{-3}$ to 1500 $\mu g\ m^{-3}$ and the level of SO_2 rose from about 0.15 ppm to 0.75 ppm due to a temperature inversion

---

[7] Particulates smaller than 10 μm (termed PM_{10} particulates) are not filtered by the lungs' cilia and can enter the lungs and cause damage. Particles smaller than 2.5 micrometers ($PM_{2.5}$) are the most dangerous because they can penetrate all the way into the alveoli of the lungs and be absorbed into the blood.

Table 1.1. Pollutants emitted per kWh electricity production, including fuel cycle and construction of the plant (Koch 2001, IPCC 2014, EPA 2014b). The ranges reflect the variability of fuel sources (for example low sulfur versus high sulfur coal) and the variability of manufacturing processes (in the case of wind and solar).

Source	CO_2 (g kWh^{-1})	SO_2 (mg kWh^{-1})	NO (mg kWh^{-1})	NMVOC (mg kWh^{-1})	Particulates (mg kWh^{-1})	Methane (g CO_2 equivalent per kWh)
Hydropower	2–48	5–60	3–42	0	5	88
Nuclear	2–59	3–50	2–100	0	2	0
Wind	7–124	21–87	14–50	0	5–35	0
Solar	13–731	24–490	16–340	70	12–190	0
Biomass	15–101	12–140	701–1950	0	217–320	0
Natural gas (combined cycle)	385–511	4–15 000	13–1500	72–164	1–10	91
Coal (modern)	790–1182	700–32 321	700–5273	18–29	30–663	47

which trapped pollutants in the London area. In lock step with this tripling in emissions the number of deaths increased from around 250 per day to 900 per day during this one week period. Similar but less severe events occurred in London 1956, 1957 and 1960. This and comparable episodes occurring in other cities have made a compelling argument for the connection between air pollution and health. We now have a reasonable amount of epidemiological data related to air pollutants and consequently air standards are now mandated in many countries.

The exact exposure rate of a population to emissions from a coal fired power plant or other industrial sized contamination source will depend on the location of the population relative to the plant, but a realistic estimation for exposure to sulfur from a local coal plant might be 5.0 μg m^{-3} of sulfur dioxide per person per tonne of sulfur dioxide emitted[8]. As an example of the uncertainty of the dose response, estimations of the effect of SO_2 based on epidemiological studies varies between two excess deaths (National Academy of Sciences estimate) to 37 excess deaths (Brookhaven National Laboratory) per million persons exposed at a concentration of 1.0 μg m^{-3} per person. Suppose the amount of SO_2 produced per year by a power plant is 5.0×10^4 tonnes yr^{-1}. The lower death estimate yields $5.0 \times 10^4 \text{ tonnes yr}^{-1} \times \frac{5\ \mu\text{g m}^{-3}}{\text{tonne SO}_2\text{ yr}^{-1}} \times \frac{2 \text{ deaths}}{10^6 \times 1\ \mu\text{g m}^{-3}} = 0.5$ excess deaths per million persons per year, while the Brookhaven estimate gives 185 excess deaths per million persons per year. Once the effect of a particular dose of a pollutant is known, the effect on a given population with a known exposure can be calculated.

---

[8] This is 10^{-5} times the US EPA mandated limit of 0.5 ppm (0.5 g m^{-3}) for exposure to sulfur dioxide, a figure that has clear negative health effects.

Contact with a contaminant may come from airborne exposure, liquids in contact with the skin or ingested substances such as food additives. Often it is not possible to measure the actual exposure, possibly because it happened sometime in the past in which case extrapolations of exposure rates are made from known data. Such a calculation is based on a *stock-flow model*[9], of which there are several versions, depending on the situation. Exposure may be from a single event, the concentration of which may or may not decrease over time or, alternatively, from a substance which is continuously being emitted into and removed from a volume of air or liquid. Examples of more or less continuous exposure are ambient pollutants from nearby factories or power generation plants. Single events might include a single radioactive leak from a nuclear power plant or a one-time chemical spill. Exposures might also be periodic, for example food additives which are ingested on a daily basis or occupational exposure which occurs on a daily or seasonal basis. Two examples are given in the following, the reader is referred to the literature for more details of extended stock-flow modeling (Kammen and Hassenzal (1999) and references therein).

The equilibrium concentration in a closed space of a constantly emitted source can be calculated if the emission rate is known. Suppose a room in a manufacturing plant has an air exchange rate of once every 4 h but a chemical vapor is being emitted from an industrial process at a rate of 12.0 $\mu\text{g min}^{-1}$. At equilibrium concentration the flow into the room from the source equals the flow out of the room due to ventilation. The removal flow rate of the vapor out of the room is $\text{flow} = VCE$, where V is the volume of the room, E is the air exchange rate and C is the concentration. For a 200.0 m^3 room the concentration would be $C = \frac{\text{flow}}{VE} = 0.9\ \mu\text{g m}^{-3}$.

In many cases, however, the source flow rate is not a constant, either because the source varies or the substance decays, chemically reacts or is absorbed. If the chemical in the previous example is being removed by an air purifier, the flow out of the room is $\text{flow} = VC(E + R)$, where R is the rate of removal by the purifier. The rate of removal, R, could also be the exponential decay rate for a radioactive substance or a chemical reaction rate for a chemical reaction which is absorbing the pollutant.

For a one time exposure the rate at which a toxin disappears is often proportional to the concentration. This is true for substances introduced directly into the blood by injection and by ingestion if the substance can easily cross the stomach lining, as well as inhaled substances passing through the lung linings. As the body removes the substance the rate of removal is proportional to the amount left. If $C(t)$ is the concentration at time t then we have

$$\mathrm{d}C(t)/\mathrm{d}t = -kC(t), \tag{1.9}$$

---

[9] Stock refers to a constant amount whereas flow refers to a changing amount of some quantity. These concepts can be applied to levels of contaminates in a building, chemical substances in the blood or stocks of capital subject to flows of income or expenditures.

where $dC(t)/dt$ is the rate of change and k is the rate constant. The negative sign indicates the concentration is decreasing. Equation (1.9) is the differential equation for exponential decay and has solutions

$$C(t) = C(0)e^{-kt}, \tag{1.10}$$

where $C(0)$ is the initial concentration. For radioactive decay the decay constant is related to the half-life of the decay process, $k = \ln(2/t_{1/2})$, where $t_{1/2}$ is the half-life[10].

The length of time for a particular exposure is calculated the same way for an undesired contact with a contaminant as for the desired exposure to a medicine. In both cases the length of exposure depends on the rate of elimination of the substance from the body. Suppose a person is injected with 1.2 g of a substance which is eliminated from the body at an exponential decay rate of $k = -0.6\ \mathrm{h}^{-1}$. An adult human body has about 5.0 l of blood resulting in an initial concentration of $C(0)=$ 1.2 g/5.0 l. Suppose that an effective dose (good in the case of medicine, bad in the case of a poison) is 0.04 g l^{-1} in the blood. Equation (1.10) can be solved for the time when $C(t)$ drops below 0.04 g l^{-1} which is 3.0 h of exposure at a concentration above the effective dose.

The assumption that the risk (or effect) is directly proportional to the dose (or cause) is known as the *linear default model*. The death rate due to lung cancer among miners exposed to radon (a short half-life radioactive daughter product of uranium decay) is known to be between 200 and 600 deaths per million working-level month (WLmo) and can be used to extrapolate risk to the general population. Working-level month is a pre-determined exposure for a given amount of time. For radon the exposure is 100 pCi l^{-1} of radon for 170 h a month as determined by the US EPA. The average home level of radon in the US is 1.5 pCi l^{-1} [11] and if we assume a person spends three fourths of their time at home this constitutes $0.75 \times 1.5\ \mathrm{pCi\ l^{-1}} \times 365\ \mathrm{day\ yr^{-1}} \times 24\ \mathrm{h\ day^{-1}} \times 1\ \mathrm{WLmo}/(170\ \mathrm{h\ mo^{-1}} \times 100\ \mathrm{pCi\ l^{-1}}) = 0.58\ \mathrm{WLmo\ yr^{-1}}$ for the average exposure in the US. Multiplying by a lifetime of 72 years gives 42 WLmo in a lifetime. The miner cancer rate of 200 per million working-level months extrapolates to $200 \times 42\ \mathrm{WLmo}/10^6\ \mathrm{WLmo} = 0.0084$ or a 0.84% chance of lung cancer for a North American exposed at the average rate for a lifetime. For the higher death rate figure, 600 deaths per million WLmo, the chance is 2.5%. For an estimated population of 306 million we have $306 \times 10^6 \times 0.58\ \mathrm{WLmo} \times 200/10^6\ \mathrm{WLmo} = 35\,500$ deaths per year, close to the number of people killed in traffic accidents per year in the US. The actual figure estimated by the US EPA is slightly lower, around 21 000 deaths per year. This is because the average level of 1.5 pCi l^{-1} does not account for the fact that many houses have no radon whereas a small number of houses have a disproportionally high concentration of radon, so the at-risk portion of the population is actually much smaller than the entire population.

---

[10] Half-life, discussed in more detail in chapter five, is the time taken for half of a radioactive substance to decay and is known for most radioactive substances.

[11] pCi l^{-1} = pico-Curie per liter; see chapter 5.

There are many examples of epidemiological data establishing linear connections between varying exposure and a given health risk. In the case of smoking, cancer rates among non-smokers, smokers with different smoking habits, and non-smokers exposed to second hand smoke clearly show an increase in risk directly proportional to the increase in exposure. The association between colon cancer and a high fat diet came from studies involving immigrants coming to the US. Before changing their diets to match those of people in the US these immigrants had lower rates of colon cancer but the rates changed to match those of US citizens as their diets changed to match those of North Americans in proportion to the extent of dietary substitutions.

In the US the Food and Drug Administration (FDA) or the EPA require animal testing for new chemicals or food additives introduced to the public. New medical treatments are similarly regulated. These studies are quite expensive some costing as much as five million dollars for a study with a few hundred rats (Kammen and Hassenzal 1999). Because of these costs newly produced chemicals not used for medicinal purposes are sometimes not tested if there is little reason to think they might be toxic or if the exposure is thought to be slight. If the epidemiological data begin to indicate a problem after the chemical has been on the market for a while a regulatory agent such as the EPA or FDA may require the manufacturer to perform animal tests.

Special bred animals, typically rodents, are used for most animal testing. Genetically identical animals, selected to be sensitive to carcinogens are preferred in order to make any response more obvious. The procedure is to give identical groups of animals varying amounts of the toxin under study and after some period of time the animals are sacrificed (i.e. killed) and examined for tumors to see if exposure correlates to increases in cancer rates. A second kind of test is to give different groups of animals different amounts of a toxin and count how many die prematurely during some set period of time. These results are reported as the LD_{50} dose, the dose that kills 50% of the test animals in some standard time interval, possibly the animals' normal, unexposed lifetime. LD_{10} values, which give the dose that kills 10% of the animals, are also often reported. In some cases toxicity studies can be performed on biological tissue or cell samples rather than live animals, however, because of the complex interaction of various systems in a living organism, animal studies are generally considered to be superior to tissue studies. A drawback of animal study results is the fact that there is no guarantee that they will extrapolate well to humans, particularly for chemicals that target specific tissues or organs which may not function comparably in humans and animals.

The concept of a linear relation between dose and effect must be modified for information from animal testing because the effects on small animals cannot be assumed to extrapolate in a strictly linear way to humans with larger body mass. A similar problem is encountered when determining appropriate dosage of medicine for children when the effective dose for an adult is known. The average daily dosage, d_h for a human that has the same effect for a given animal dose, d_a is $d_\mathrm{h} = \mathrm{sf} \times d_\mathrm{a}$, where the scale factor, sf, is given by $\mathrm{sf} = (W_\mathrm{h}/W_\mathrm{a})^{1-b}$. Here W_h is the weight of the human and W_a is the weight of the animal, and b is usually taken to be approximately 0.75, a value based on historical data and used by the US EPA. Suppose the rats in an exposure trial averaged 0.5 kg and had a normal lifetime of 2.5 yr. Further suppose a dose of 30.0 mg kg^{-1} per day for a group of 50 rats resulted in ten tumors (a probability for tumor development of 0.20). Body weight is proportional to mass so for a 65 kg human we have $d_\text{h} = (W_\text{h}/W_\text{a})^{1-b} \times d_\text{a} = (65.0/0.5)^{0.25} \times 30.0\ \text{kg}^{-1} =$ 101.3 mg kg^{-1} per day for the human. Were a group of humans to be exposed to this daily dose of the drug we would expect a 20% (0.20 probability) incidence of tumors. The total dosage per day needed to produce this effect in a 65 kg person would be 101.3 mg $\text{kg}^{-1} \times$ 65 kg = 6.6 g (adapted from Kammen and Hassenzal (1999)).

In many risk analyses we want to evaluate whether the results of two different datasets are significantly different in a statistical sense. For example, suppose one group of test animals is exposed to a toxin and develops cancer at a rate of 20% while an identical group is not exposed to the toxin but ends up with a 10% rate of the same type cancer. Can we conclude that the toxin causes cancer and if so, with what confidence? A full discussion of the statistics involved in making these kinds of decisions is beyond the scope of this text but a simple method of hypothesis testing in the case of only two possible outcomes (tumor or no tumor in this example) is the *z-test* which is based on the standard deviation of the binomial distribution. Other tests available for distinguishing possible cause and effect in trials where the data are not obviously conclusive include the *chi-squared test*, *Student's t-test* and the *f-test*. These kinds of statistical tests can determine the likelihood that exposure to a given contaminant is implicated in a disease outcome as a percentage chance that the two are linked. The reader should consult more advanced texts for further explanation of statistical methods (Bevinton and Robinson 1992, Lyons 1991, StatSoft 2013).

There is a great temptation for the general population to demand a zero exposure rate to any hazardous substance, but there are at least two reasons not to do this. For most substances there is a threshold dose below which the effects cannot be discerned from other causes of death in the general population. It is probably impossible and not necessary to insist that public drinking water have completely undetectable amounts of arsenic (Wilson and Crouch 2001). A more reasonable approach is to try to reduce the risk to a reasonable level, for example one in 100 million. At this rate there would be fewer than three cases of cancer caused by arsenic in the entire US per year which would be very difficult to distinguish from other cancer causes. In effect lowering the arsenic exposure risk to one in 100 million lowers it below the background of other causes of death and we can no longer be sure that someone with cancer got it from their exposure to arsenic. The cost of reducing detectable arsenic levels to zero is also prohibitive.

For exposure to many substances the effects (harmful or beneficial) may not be linear. It is known, for example, that the human body continues to perform a certain amount of repair to DNA and other important biological systems. Figure 1.9 shows the linear dose model and several alternatives. In the figure a positive response may indicate a beneficial or a detrimental effect depending on whether the substance is a medical drug or toxin. Alcohol consumption is an example of *hormesis* where a positive dose response is considered to be detrimental; drinking alcohol has adverse risks (cancer, accidental death) if consumed at doses of more than three

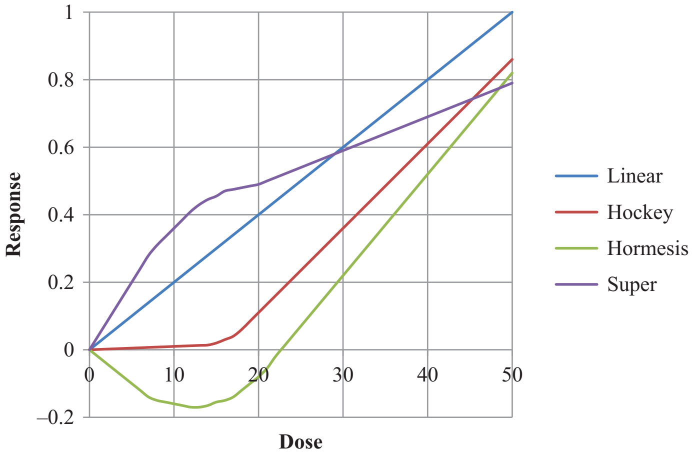


Figure 1.9. Possible dose response curves in arbitrary units where 'response' may be considered to be bad (in the event of toxins) or good (if the substance is a medication). From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

drinks per day. However it is also the case that people drinking less than one drink per day have higher death rates than those drinking one to two drinks per day. The death rate associated with alcohol is lowest for men drinking one and a half drinks per day and for women drinking one drink a day, where one drink is defined to be one 12 oz beer, one 5 oz glass of wine or one 1.5 oz shot of whisky. The probable reason for this effect is that alcohol increases cancer risk and risk of accidents but decreases risk from heart disease, and these two opposing effects intersect at around one drink per day.

In the *threshold* or hockey stick model there is little or no reaction until a threshold dose, sometimes called the *maximum tolerated dose*, is reached (arround 15 dose units in figure 1.9). Some types of carcinogens seem to fit this model; because of repair mechanisms in the body low exposure rates have no apparent effect over the lifetime of the subject. Tolerance (to good or bad drugs) due to constant exposure fits the *super dose response* curve; initially the drug causes a large response but taking more of the drug does not produce a proportional increase in response. Differentiating between the various possible response curves requires sufficient data, either from animal studies or historical data. For example, if only a few data points at doses between 45 and 50 units in figure 1.9 were known, it would not be possible to distinguish between the four dose curves.

Synergistic effects for exposure to two or more potentially harmful agents simultaneously are very difficult to establish, particularly if there is a time delay between exposure and effect. The interaction between radon and cigarette smoking, for example, required a great deal of research since both cause lung cancer and until recently the amount of radon exposure for various groups was not well established. It is now known from epidemiological data from miners who often recieve larger doses of radon, that smoking and radon together cause a greater risk than simply multiplying the risk of each.

Projects

1. Pick one or several countries in the Food and Agricultural Organization of the United Nations www.fao.org/home/en/ database. Analyze water, food security, fisheries and agricultural resources by year for this country or region. Compare the results of analysis between a developed country and a developing country.
2. Investigate the problem of overdrawing from aquifers or some other water resource somewhere in the world. Include a discussion of the local environmental effects and legal aspects of any conflicts about who owns a water resource, who gets to decide how a resource is used, etc. How does this relate to green, blue, grey or virtual water?
3. Write a report on desalination of sea water. Find out what methods are being used and compare costs in energy and dollars. Which are cost effective compared to today's cost of water?
4. Check mortality statistics for the US at the Centers for Disease Control and Prevention (CDCP 2016). What are the leading causes of deaths? Write a report explaining these and why they have changed over time.
5. Verify the claims in the text about population, fertility, mortality, aging, migration and urbanization for a country or a region using the United Nations Population Division www.un.org/en/development/desa/population/. In particular, show that the increase in population is exponential for some time span and find the exponential increase rate. (Hint: A quantity $N(t)$ is said to be exponential if, at time t we have $N(t)=N(0)e^{rt}$, where $N(0)$ is the original quantity at $t = 0$ and r is the (constant) growth rate. Taking a log of this equation gives $\ln[N(t)] = \ln[N(0)] + rt$, which is a straight line with slope r (it is much easier to plot a straight line than an exponential curve!).)
6. Use the hint in the previous question to show that some other commodity such as cell phones, refrigerators, pet snakes, etc, has undergone an exponential increase that matches or exceeds the exponential increase in the population for some period of time. Discuss the factors involved.
7. In figure 1.3 the global fish catch is seen to level off after 1990. Find out if this is because ocean fisheries around the world are collapsing due to overfishing or if it is because of an increase in fish farming or from some other effect, and write an essay explaining your conclusions.
8. Write a comprehensive report on bio-remediation for various pollutants or some other novel technique of removing pollution from soil, air or water.
9. For some state or region in the US, analyze the trends in municipal waste, air and water pollution, toxic chemicals and land recovery using US EPA www3.epa.gov/data. What is being done (legally and physical mechanisms) to reduce environmental pollution. Are some strategies more successful than others? Why? Discuss how these factors interact in your analysis.
10. Go to Gapminder World; Health and Wealth of Nations www.gapminder.org and examine some of the statistics in the animated chart. How does energy use compare to per capita GDP over time? How about fertility rates

and GDP per capita? Describe any other interesting correlations you find. Are there global trends for different countries over time? Explain.

11. Write a report on the details of the IPAT equation and compare it to any alternatives you might find.
12. Go to NASA's Sulfur Dioxide monitoring web page https://so2.gsfc.nasa.gov/index.html and report on current pollution trends. Many of these maps are animated so that you can watch how pollution circulates around the globe. Report on any interesting trends for sulfur dioxide and other global pollutants.
13. Report on the latest information about CFCs, their replacements (which are now being replaced by even newer compounds) and the status of the ozone hole. Why are the original replacements now being phased out?
14. Find out how many of the following items are discarded each year (either globally or in some specific country for which there are good statistics): cell phones, computers, televisions, refrigerators, and some other item such as ballpoint pens, toys, etc. Are these increases exponential? Link these figures to recycling efforts of these objects.

References

Bevinton P R and Robinson D K 1992 *Data Reduction and Error Analysis for the Physical Sciences* (New York: McGraw-Hill)

Biello D 2013 The world's 10 most polluted places *Scientific American online slide show* www.scientificamerican.com/article/10-most-polluted-places-in-the-world1/ (Accessed: 6 August 2016)

CDCP (Centers for Disease Control and Prevention) 2016 http://blogs.cdc.gov/nchs-data-visualization/deaths-in-the-us/ (Accessed: 2 August 2016)

Collier P 2007 *The Bottom Billion* (Oxford: Oxford University Press)

Commoner B 1972 The environmental cost of economic growth *Population, Resources and the Environment* (Washington, DC: Government Printing Office) pp 339–63

EPA (Environmental Protection Agency) 2014a Municipal solid waste generation, recycling, and disposal in the United States: tables and figures for 2012 www.epa.gov/sites/production/files/2015-09/documents/2012_msw_dat_tbls.pdf (Accessed: 2 August 2016)

EPA (Environmental Protection Agency) 2014b National Emissions Inventory (NEI) data www.epa.gov/air-emissions-inventories/2014-national-emissions-inventory-nei-data

EPA (Environmental Protection Agency) 2016 Acid Rain Program www.epa.gov/acidrain (Accessed: 5 August 2016)

Falchil F *et al* 2016 The new world atlas of artificial night sky brightness *Sci. Adv* **2** e1600377

FAO (Food and Agricultural Organization of the United Nations) 2016 http://faostat3.fao.org/download/Q/QA/E (Accessed: 2 August 2016)

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Hafemeister D 2007 *Physics of Societal Issues* (New York: Springer)

IPCC (Intergovernmental Panel on Climate Change) 2014 Climate Change 2014; the Mitigation of Climate Change (Cambridge: Cambridge University Press) table A.III.2

Kammen D M and Hassenzal D M 1999 *Should We Risk It? Exploring Environmental, Health, and Technological Problem Solving* (Princeton, NJ: Princeton University Press)

Koch F H 2001 Hydropower-internalized costs and externalized benefits *Proc. Externalities and Energy Policy: The Life Cycle Analysis Approach Workshop (Paris, France, 15–16 November)* www.oecd-nea.org/ndd/reports/2002/nea3676-externalities.pdf

Lelieveld J, Evans J S, Fnais M, Giannadaki D and Pozzer A 2015 The contribution of outdoor air pollution sources on premature mortality on a global scale *Nature* **525** 367

Litterbase 2017 http://litterbase.awi.de/ (Accessed: 26 April 2017)

Lyons L 1991 *A Practical Guide to Data Analysis for Physical Science Students* (Cambridge: Cambridge University Press)

McNeill J R 2000 *Something New Under the Sun; An Environmental History of the Twentieth-Century World* (New York: Norton)

Mekonnen M M and Hoekstra A Y 2011 National water footprint accounts: the green, blue and grey water footprint of production and consumption *UNESCO Institute for Water Education* http://waterfootprint.org/media/downloads/Report50-NationalWaterFootprints-Vol1.pdf (Accessed: 2 August 2016)

NASA (National Aeronautics and Space Administration) 2016 Goddard Space Flight Center, Atmospheric Chemistry and Dynamics Laboratory http://so2.gsfc.nasa.gov/index.html (Accessed: 5 August 2016)

National Water Footprint 2016 http://waterfootprint.org/en/resources/interactive-tools/national-water-footprint-explorer/ (Accessed: 3 August 2016)

Nielson J 2002 The killer fog of '52; thousands died as poisonous air smothered London *National Public Radio* www.npr.org/templates/story/story.php?storyId=873954 (Accessed: 2 August 2016)

Norberg J 2016 *Progress: Ten Reasons to Look Forward to the Future* Oneworld

Nriagu J O 1996 A history of global metal production *Science* **272** 223

Ramanathan V *et al* 2001 Indian Ocean experiment: an integrated analysis of the climate forcing and effects of the great Indo-Asian haze *J. Geophys. Res.* **106** 371–98

Rich C and Longcore T 2005 *Ecological Consequences of Artificial Night Lighting* (Washington, DC: Island)

Shargorodsky J, Curhan S G, Curhan G C and Eavey R 2010 Change in prevalence in hearing loss among US adolescents *J. Am. Med. Assoc* **304** 772–8

Smil V 2015 *Harvesting the Biosphere: What We have Taken from Nature* (Boston, MA: MIT press)

StatSoft 2013 *Electronic Statistics Textbook* www.statsoft.com/Textbook

United Nations 2013 *Statistical Yearbook* http://unstats.un.org/unsd/publications/statistical-yearbook (Accessed: 3 August 2016)

United Nations Population Division 2015 *World Population Prospects: Key Findings and Advanced Tables* https://esa.un.org/unpd/wpp/Publications/Files/Key_Findings_WPP_2015.pdf (Accessed: 2 August 2016)

United Nations Population Division 2016 https://esa.un.org/unpd/wpp/ (Accessed: 11 July 2016)

van Sebille E, Wilcox C, Lebreton L, Maximenko N, Hardesty B D, van Franeker J A, Eriksen M, Siegel D, Galgani F and Law K L 2015 A global inventory of small floating plastic debris *Environ. Res. Lett.* **10** 124006

Weeks J R 2016 *Population, An Introduction to Concepts and Issues* (Boston MA: Cengage Learning)

Wilson R and Crouch E A C 2001 *Risk-Benefit Analysis* (Cambridge, MA: Harvard University Press)

IOP Concise Physics

Physics and the Environment

Kyle Forinash III

Chapter 2

Energy conservation and the first law of thermodynamics

Energy is a fundamental property of the physical world that can be transferred between objects to cause physical change such as mechanical work or temperature changes. The *first law of thermodynamics* is equivalent to the *law of conservation of energy*: energy cannot be created or destroyed; the total amount of energy in the Universe is fixed. Energy can be transformed from one form to another or transferred from one place to another but the total energy must remain unchanged. In a *closed mechanical system* no energy is added or subtracted from the system while a *closed thermodynamic system* conserves the physical material in the system and known quantities of energy are allowed to enter and exit the system. *Heat* is defined as an energy exchange as the result of a temperature difference between two objects (temperature is defined in the following). The behavior of many environmental and human designed systems can only be understood by examining heat energy exchanges and conversions related to the functioning of those systems, which is the topic of this chapter.

2.1 The first law of thermodynamics

The quantities *work*, W, *heat*, Q, and changes in *internal energy*, ΔU, are the most useful for describing a thermodynamic system. If energy is added to the system and work is done on the system Q and W are positive and have the opposite signs for heat leaving the system and work done by the system. These quantities are typically measured in units of joules (J) or calories (cal) where 1.000 J = 0.2389 cal. The simplest form of the first law of thermodynamics for processes where only energy flows into or out of the system is

$$\Delta U = Q + W. \tag{2.1}$$

doi:10.1088/978-1-6817-4493-3ch2

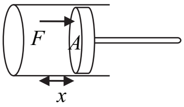


Figure 2.1. A piston with area A moving through a distance dx will sweep out a volume dV. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

The symbol Q, often referred to colloquially as heat flow, refers to a transfer of energy as the result of a temperature difference. One cannot talk about the heat in an object and it is incorrect, strictly speaking, to talk about heat flow, nevertheless, two objects at different temperatures in contact with each other will exchange energy until their temperatures are the same (a process sometimes referred to as the *zeroth law of thermodynamics*). The energy exchange must occur by some combination of only three possible mechanisms, convection, conduction or radiation, which are discussed below. Evaporation also transfers energy to or from an object, however, in this case the system is not closed and a mass transfer is involved.

Work is defined in mechanics as $W = \int_{x_\mathrm{f}}^{x_\mathrm{i}} \vec{F} \cdot \mathrm{d}\vec{s}$ where $\vec{F}$ is a force which acts over a distance between some initial location, x_i, and a final location, x_f, along the direction $\mathrm{d}\vec{s}$. In many thermodynamic applications we are interested in work done by an expanding or contracting gas. We may modify the mechanical definition of work for a gas or fluid pushing a piston inside a cylinder by using the definition of *pressure*, $P = F/A$ and realizing the gas will push the front surface area of the piston through a volume, $A\mathrm{d}x = \mathrm{d}V$ (figure 2.1). This results in the thermodynamic definition of work being

$$W = \int_{V_\mathrm{i}}^{V_\mathrm{f}} P\,\mathrm{d}V. \tag{2.2}$$

Pressure is measured in pascals where 1 Pa = 1 N m^{-2}.[1] Other common units of pressure are pounds per square inch (PSI), torrs, atmospheres, millimeters of mercury (mmHg) and bar or millibar (mbar).

Internal energy, U, is the sum of all of the types of energy the individual molecules may have, including rotational, vibrational and kinetic energy at the microscopic level and energy absorbed or released by chemical reactions between molecules. As a result, energy related to temperature changes (*heat capacity*[2]), phase changes (*latent heats*[3]) and chemical reactions such as burning are included in the internal energy for real substances. An *ideal gas* is defined to be one where the molecules are non-interacting point particles without rotational, vibrational or other types of internal

---

[1] On a microscopic level pressure represents the effect of billions of molecules of the gas applying a force to a side of a container striking the area in a random walk.

[2] Specific heat capacity is defined by the ratio of heat added, Q, to the temperature increase for a given mass, m: $C = Q/m\Delta T$. For materials which change volume under pressure there are two empirical values available, one for constant pressure, C_P, the other if the volume is kept constant, C_V.

[3] The latent heat of fusion, L_f, is the energy needed to change a mass of a substance from a solid to a liquid *without* changing its temperature: $L_f = Q/m$. The latent heat of vaporization, L_v, gives the energy needed to turn a sample of mass m from a liquid to a solid *without* changing the temperature: $L_v = Q/m$.

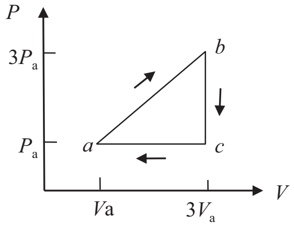


Figure 2.2. The P–V diagram for a theoretical heat engine. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

energy. The definition precludes an ideal gas from ever condensing into a liquid or solid because there are no intermolecular forces to hold the molecules together.

Practical applications of energy conversions into forms that are useful to humans include *heat engines* which are defined to be any device that uses the heat flow due to a temperature difference to provide mechanical work. The modern gasoline or diesel engine is an example of a heat engine burning gasoline to form a hot gas which expands to do mechanical work. Useful heat engines typically involve *cyclic processes* whereby a system absorbs a heat flow to do work and returns to the initial state so that the net change in internal energy at the end of a cycle is zero.

The pressure versus volume graph of a theoretical heat engine is shown in figure 2.2. In this example gas increases its pressure and volume (due to added heat), expanding along path $a \to b$. Along path $b \to c$ heat is expelled and the pressure drops while the volume remains constant. From $c \to a$ the pressure remains constant while the gas contracts and the system returns to its original state with $\Delta U_{\text{cycle}} = 0$. From the definition in equation (2.2) we know that the area enclosed in the curve shown is the work done by the system.

A P–V diagram allows a calculation of the work done by a heat flow through a theoretical system if the heat flow, pressure and volume changes are known. For example if $Q_{a\to b} = 500$ J of heat is known to be absorbed and $Q_{b\to c} = 250$ J of heat is expelled, for $P_{\text{a}} = 100$ kPa and $V_{\text{a}} = 1$ l we have $W_{a\to b} = \frac{1}{2}(200\ \text{kPa} \times 2.00\ \text{l}) + (100\ \text{kPa} \times 2.00\ \text{l}) = 400\ \text{J}$, $W_{b\to c} = 0$ and $W_{c\to a} = -100\ \text{kPa} \times 2.00\ \text{l} = -200\ \text{J}$ (work done on the system to reduce the volume is negative). The total work done is then $W_{\text{cycle}} = 400\ \text{J} - 200\ \text{J} = 200\ \text{J}$ (the area inside the curve). Using $\Delta U_{\text{cycle}} = 0$ and the first law we have $0 = W_{\text{cycle}} + Q_{\text{cycle}}$ or $0 = 200\ \text{J} + Q_{a\to b} - Q_{b\to c} + Q_{c\to a}$ from which we have $Q_{c\to a} = -450$ J.

The first law says that the total energy of a closed system would remain fixed if there are no energy exchanges with the surroundings. If we add heat energy to a substance via a temperature difference and do work on it (change the volume), the internal energy will change by an amount equal to the added energy. We can retrieve that energy but only at the expense of the internal energy. A claim that a cyclic process where $\Delta U_{\text{cycle}} = 0$ can output more energy than was originally added is called a *perpetual motion device of the first kind* and would break the first law of thermodynamics and so is not achievable. The plethora of proposals found on the

Internet such as using water as a fuel for automobiles or using some combination of magnets or flywheels to generate unlimited amounts of ‘free energy’ are not possible.

2.2 Efficiency

In the previous example of a cyclic process it should be noted that not all of the input heat was converted into work; some heat had to be expelled to return the system to its initial state. This is a general feature of all heat engines and is the result of the *second law of thermodynamics*, which states that no cyclic process can convert all the input energy into useful work. The second law will be discussed in more detail in the next chapter. *Efficiency* measures this loss of potentially useful energy in a process and is defined to be the ratio of the energy benefit to the energy cost:

$$\eta = \frac{\text{energy benefit}}{\text{energy cost}} = W_{\text{cycle}}/Q_{\text{H}}, \tag{2.3}$$

where Q_{H} is the heat flow into the process from a hot reservoir (e.g. burning gasoline). The first law limits the efficiency of a heat engines to 100%—you cannot get more energy out of a cyclic process than input energy and the second law says the actual efficiency unavoidably will be much less than 100%. Using this definition the efficiency for the heat engine in the example above is $W_{\text{cycle}}/Q_{\text{H}} = \frac{200\text{ J}}{500\text{ J}} \times 100\% = 40\%$. Burning natural gas for heat is very efficient since the benefit is thermal energy; gas furnaces can be more than 95% efficient since nearly all the output energy is useful as heat (some energy is lost in the exhaust gasses). However converting thermal energy into mechanical energy with a heat engine has much lower efficiency; gasoline engines are less than 25% efficient in turning the chemical energy in gasoline into kinetic energy of the car[4].

Efficiencies for several energy conversion processes are shown in table 2.1. The efficiency of all energy conversion processes is limited by the second law of thermodynamics, but it should be clear from the table that some processes are inherently more efficient than others. Electrical conversions (which include motors and batteries), are in general much more efficient than heat engines (which include internal combustion processes such as gasoline and diesel engines). There are several ways to convert electricity into light but none of these processes is particularly efficient. The table lists lighting efficiencies in energy terms but many other sources list *efficacy* which is the lumens output (light visible to the human eye) per watt; see the discussion below. Plants are nearly 90% efficient at capturing light but very inefficient at turning this energy into a stored form which would be useful for other purposes. As we will see later, this makes fuels derived from plants much less attractive than is often assumed (see chapter 6).

In most technical applications of the use of energy there are a series of linked processes, where several energy conversions take place, each with a known

---

[4] It is estimated that end use power demands would decrease by 39% if all heat engine applications were replaced by electrical processes (Jacobsen *et al* 2015).

Table 2.1. A few energy conversion processes and their efficiencies after (Smil 1999).

Conversion process	Type of process	Typical efficiency (%)	Laboratory or theoretical maximum efficiency (%)
Batteries	Chemical to electrical		
Lithium ion		80–90	
Nickel cadmium		66	
Lead storage		45	
Electric motors	Electrical to mechanical		
Large (10 000 W)		92	
Small (1000 W)		80	95
Internal combustion	Chemical to mechanical		
Diesel		50	60
Gasoline		20	37
Steam		25	50
Gas turbine	Chemical to electrical	60	70
Plants	Radiant to chemical		
Light capture		95 (visible only)	
Total efficiency (most plants)		1–2	6–8 (see Zhu *et al* 2008 and Kromdijk *et al* 2016)
Sugar cane		1.2	
PV cells	Radiant to electrical	20	46
Thermoelectric	Thermal to electrical	6–8	
Solar thermal	Radiant to thermal	40	
Fuel cells	Chemical to electrical		
Proton exchange		30–50	60
Methanol		25–40	83
Hydroelectric	Gravitational to electrical	75	
Lighting	Electrical or chemical to radiant		
Candle		<1	
Incandescent		3	5
Fluorescent		10–12	15
LED		4–10	22
Sodium discharge		22–27	

efficiency. The total efficiency is the result of multiplying the efficiency of each step. Consider the electricity use for an electric fan in a house. This process probably starts with a coal fired plant which makes steam to turn a turbine which turns a generator to make electricity. This process has an efficiency of around

32% for turning the chemical energy in coal into the energy carried by the electric current[5]. Transmission efficiency of electricity along high voltage lines is around 90% and the motor in the electric fan might have an efficiency as high as 80% in turning electricity into the mechanical motion of the blades. The total efficiency of this process is then $\eta = 0.32 \times 0.90 \times 0.80 = 0.23$ or 23%. The limiting efficiency in a series of energy conversions is the step with the lowest efficiency which is the electricity generation in this example. This consideration makes it clear that burning natural gas for home heating is much more efficient from an energy perspective than using natural gas to make electricity (55% efficient) to be used for home heating (although cost often is more of a decisive factor when making a decision about home heating).

Most heat engine applications today shed the non-beneficial energy as heat. Electricity generation requires either cooling towers or a lake or river to dispose of the 'waste' heat (which constitutes 68% of the input energy for coal and 45% for modern gas super-efficient turbines). Car engines have a radiator to transfer the excess heat to the environment (more than 75% of the original energy embodied in the gasoline). This waste heat can and has been used for other purposes and this process is called *cogeneration*[6]. Combined heat and power plants are purposely designed to use the waste heat from electricity generation using fossil fuels. For example, modern paper manufacturing requires the harvested wood chips be cooked for several hours at 180 °C in a chemical bath to break down the cellulose in the wood (the Kraft process). The heat from this process is sufficiently high to also generate steam for electricity production to run the paper mill plant.

It has been known since the early 1800s that certain metals in contact with each other would produce electric current if heat was applied. This has been used to make thermocouples, which are basically solid state thermometers, and to convert waste heat directly into electric current via the *thermoelectric* effect. Although the materials used so far to make thermoelectric generators involve expensive rare earth metals, the process is being investigated as a way to capture waste heat from car engines and conventional electric power plants. Efficiencies are currently less than 10%, but the input energy is energy that would otherwise be wasted.

2.3 Energy transfer

The macroscopic quantity *temperature* is defined empirically using the volume of various gases as heat is removed. Absolute zero kelvin (0 K = −273.15 °C), is defined as the temperature at which the volume of a gas would reach zero when cooled (graphs of the volume of various gases have different slopes that all extrapolate to

---

[5] Newer design ultra-supercritical coal plants use finely powdered coal to achieve efficiencies up to 43%.

[6] In the early 1900s as much as 30% of the waste heat from electricity generation in the US was used for heating homes and businesses but this gradually changed as power plants were moved further from residential areas due to pollution.

zero at −273.15 °C as they are cooled). Statistical mechanics tells us that this macroscopic quantity is related to the average kinetic energy of a gas:

$$\frac{3}{2}k_{\mathrm{B}}T = \left\langle \frac{1}{2}mv^2 \right\rangle, \tag{2.4}$$

where $k_{\mathrm{B}} = 1.38 \times 10^{-23}$ J K^{-1} is Boltzmann's constant and the brackets denote average[7]. Although this expression is not quite right because it predicts zero kinetic energy at zero temperature (which is not the case as a result of quantum mechanics) it is a useful relation for connecting macroscopic measured quantities with microscopic theoretical quantities.

As stated above, heat is defined to be energy transferred from one object to another due to a temperature difference. There are four such mechanisms of transferring energy and it is important to realize that these are the *only* available means of transporting thermal energy between objects at different temperatures. One, evaporation, includes a transfer of mass while the other three, convection, conduction and radiation, do not.

Heat transfer by *conduction* involves the direct contact of two substances where the molecules have different average kinetic energies. The faster moving molecules at the surface of the warmer sample will bump into slower moving molecules at the surface of the cooler sample, transferring energy in a *diffusion* process[8]. Most applications involve energy transfer through a slab of material, such as an insulating wall, the skin or a fur coat. An empirical equation for the energy transfer per time is

$$\mathrm{d}Q/\mathrm{d}t = kA\Delta T/L, \tag{2.5}$$

where $\mathrm{d}Q/\mathrm{d}t$ is the rate of energy transfer or *power* (in J s^{-1} or watts) through a slab of material of area A and thickness L. The temperature difference from one side of the material to the other is ΔT and the parameter k is the thermal conductivity, a constant which depends on the insulating material and given in units of W K^{-1}m^{-1}. Representative values of thermal conductivity for common substances are given in table 2.2.

Decreasing the rate of heat transfer through a wall or barrier involves choosing a material with a lower conductivity, increasing the thickness of the material or decreasing the area of the material. It is not practical to change the material or thickness of a window beyond a certain point, so many Scandinavian countries regulate the allowed size of the window in home and business construction to reduce heat loss. Because the rate of heat transfer depends on the temperature difference across the slab, better window insulation is less important in the summer or in warm climates than in cooler climates and in winter when the difference between inside and outside temperatures is larger.

---

[7] Einstein is credited with discovering the principle embodied in equation (2.4) in his explanation of Brownian motion. Very small particles seen in a microscope have a random motion due to bombardment by the thermal motion of even smaller atoms and molecules.

[8] In this case the difference in average kinetic energy diffuses through the bulk material. Other examples of diffusion involve the movement of other types of atoms or molecules moving through the bulk material. Both types obey similar equations and so are termed diffusion.

Table 2.2. Thermal conductivities of common construction materials.

Material	Thermal conductivity, k, in W $K^{-1}m^{-1}$
Copper	401
Aluminum	237
Iron	79.5
Concrete	0.9–1.3
Glass	0.8
Water	0.6
Wood	0.05–0.36
Fiberglass insulation	0.04
Air (still)	0.02
Styrofoam	0.01

Some insulating materials can have different conductivities depending on humidity. For example dry clothing typically has a conductivity of 0.045 W $m^{-1}K^{-1}$ but the same clothing soaked with water can change the conductivity to 0.65 W $m^{-1}K^{-1}$. This changes the rate of heat loss by more than a factor of 10, which makes being outdoors in cool weather with wet clothing miserable. We can estimate the rate of energy loss for this case by assuming a cylindrical human, 1.50 m tall and 0.30 m in diameter with a 1.50 cm layer of clothing. The surface area of this cylinder will be 1.55 m^2 and we assume heat flow is uniform over the surface. If the person's skin temperature is 34 °C and the outdoor temperature is 0 °C there will be a temperature gradient of $\Delta T = 34\ ^\circ\mathrm{C} = 34$ K across the insulating layer. This results in a heat loss of 158.2 W for the dry clothing and 2285.7 W for the wet clothing.

Heat transfer by *convection*[9] involves the movement of a fluid (liquid or gas) which first gains heat from a stationary solid by conduction (a diffusion process) and then carries this energy away (an *advection* process). Most applications of convection assume the bulk material is at a higher temperature than the moving fluid. Convection processes are difficult to quantify because they depend strongly on the fluid transporting the heat, the buoyancy of this material as it changes temperature (and volume), the shape and size of the object, its initial temperature and whether the fluid is forced to move or moves only under the effects of buoyancy and gravity. It is often the case that a thin boundary layer of fluid adheres to the surface of the stationary object, adding an insulating layer. For example, window glass is a better thermal insulator than would be expected because a layer of air adheres to the surface but wind strongly affects the depth of this layer. Multiple layers (glazing) enhance this effect by trapping several air layers on each surface.

---

[9] Convection also applies to movements of fluids due to density or other differences, for example ocean currents undergo convective flow in part due to salinity differences and air can undergo convection due to humidity differences in addition to temperature. Here we are referring to heat convection.

Natural convection occurs when a fluid near an object is heated and expands, thus becoming more buoyant. The heat transfer coefficient in this case depends on the viscosity of the moving fluid, the thermal expansion coefficient, the size of the object and the temperature difference between the object and the surrounding fluid. *Laminar* flow occurs when the fluid moves smoothly past the surface. The flow can also be *turbulent* if there are swirls and eddies near the surface. The difference between these two regimes depends on the shape of the object, the viscosity of the fluid and the temperature difference.

A simplified equation where the energy transfer is proportional to the exposed surface area, A, of the object is known as Newton's law of cooling, given by

$$\mathrm{d}Q/\mathrm{d}t = hA\Delta T. \tag{2.6}$$

Here h is the heat transfer coefficient in units of W m^{-2} K^{-1}, and will be highly dependent on the particular situation. In general h is found empirically but it can be calculated for a few special cases. Monteith and Unsworth have an excellent discussion of the various factors involved and some interesting biological cases including the cooling of plant leaves, birds, insects and other animals (Monteith and Unsworth 2008).

Two different heat transfer coefficients, h, can be determined for cases of natural convection where air removes heat from cylinders and flat surfaces depending on whether the flow is laminar or turbulent. For laminar flow, which occurs when the temperature difference is not too large and for objects roughly a meter in size, the heat transfer coefficient is approximately given by

$$h \sim 1.4(\Delta T)^{1/4}, \tag{2.7}$$

where ΔT is the temperature difference between the object and the surroundings. For turbulent flow occuring more often for larger temperature differences, and larger objects we have

$$h \sim 1.3(\Delta T)^{1/3}. \tag{2.8}$$

In part the difference in powers in equations (2.7) and (2.8) (and the related equations below for forced convection) have to do with the fact that turbulent flow reduces the boundary layer which acts as an insulating layer clinging to the outside surface of the object.

In the case of *forced* convection the heat loss will depend on the speed of the air, v. For slow, laminar air movement over a flat plate we have

$$h \sim 3.9\sqrt{v} \tag{2.9}$$

and for turbulent flow over a flat plate, when air speeds are faster, we have

$$h \sim 5.7v^{0.8}. \tag{2.10}$$

For forced convection of air over a cylinder at speeds high enough to be turbulent we have

$$h \sim 4.8v^{0.6}. \tag{2.11}$$

In all of these cases the units of h will be W m^{-2} K^{-1} if v is in m s^{-1}. The empirically derived equations shown above are approximately the same for vertical and horizontal surfaces and also work reasonably well for heat transfer in the opposite direction from a hot, convecting fluid to a solid object at a lower temperature. The coefficients will be different if the fluid is more viscous, but the divisions between turbulent and laminar flow remain.

For our cylindrical human of the same dimensions as above and under the same outdoor temperature conditions we have a natural convection coefficient of $h \sim 1.4(\Delta T)^{1/4} = 3.4$ W m^{-2} K^{-1} and a convection loss of 179.2 W, but for a wind of 15 m s^{-1} (30 mph) we have $h \sim 4.8v^{0.6} = 24.4$ W m^{-2} K^{-1} and a heat loss of 1285.9 W, nearly an order of magnitude greater. There are several alternative definitions of *wind-chill factor* which constitute attempts to quantify the subjective feeling of heat loss due to forced convection at a given temperature.

In physics the term *radiation* refers to electromagnetic waves over a wide range of wavelengths. Only radiation with wavelengths below (and frequencies above) the visible spectrum (called *ionizing radiation*) is harmful to health[10]. The total electromagnetic radiation from a hot object, called *blackbody radiation*, is independent of the exact material of the object and consists of a broad range of wavelengths. The peak in this distribution of wavelengths, the wavelengths where most of the energy is emitted, is temperature dependent as we will see below. The energy transfer due to radiation from any object above absolute zero kelvin is given by the Stephan–Boltzmann equation,

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \sigma\varepsilon A T^4, \tag{2.12}$$

where A is the surface area of the radiating object, T is the temperature in kelvins and σ is Stefan–Boltzmann's constant, $\sigma = 5.669 \times 10^{-8}$ W m^{-2} K^{-4}. Some objects are more efficient at absorbing (or emitting) photons and the *emissivity*, ε, is an empirical measure of this property. The emissivity indicates the fraction of absorbed (or emitted) radiation from the surface and varies between zero (perfectly reflecting) and one (perfectly absorbing). Good emitters are equally good absorbers. The term *blackbody* comes from the fact that an object which absorbs all energy (emissivity equal to one) and reflects nothing would appear black in color.

Equation (2.12) gives the energy *emitted* per second which is the *power* emitted, $P(T) = \mathrm{d}Q/\mathrm{d}t$ by an object at temperature T. Since emission and absorption processes are equivalent for a given object, the same equation can also be used to calculate the energy *absorbed* by an object from surroundings which are at some other temperature, T_s. Every object is both absorbing and emitting at the same time, but the temperatures that determine the power emitted and absorbed are different as are the distributions of wavelengths emitted. The net energy transfer from or to the object is $P(T_\mathrm{o}) - P(T_\mathrm{s})$. For example, suppose our hypothetical cylindrical human is indoors in an ambient temperature of 20 °C (293 K). For an emissivity of 0.50 the radiated heat

---

[10] Recall that for electromagnetic waves frequency and wavelength are related by $f\lambda = c$ where c is the speed of light.

(surface temperature 34 °C or 307 K) will be 390.3 W and the absorbed heat from the room will be 323.9 W for a net heat transfer of 66.4 W to the room.

The electrons in individual atoms and molecules have discrete energy levels determined by quantum mechanics. When these electrons change energy levels they may absorb or emit only discrete amounts of energy because the energy levels are fixed (*quantized*). This emitted or absorbed energy is in the form of *photons* where the frequency of the photon is determined by $E = hf = hc/\lambda$ where E is the energy in joules or electron volts (eV), f is frequency in Hz, c is the speed of light (2.999×10^9 m s^{-1}) and h is Planck's constant, 6.63×10^{-34} Js (= 4.14×10^{-15} eVs). The *discrete spectra* of absorbed or emitted photons is a unique signature of the atom's or molecule's energy levels and can be used to identify the particular atoms or molecules in a sample. This is useful for identifying the chemical constituents of gasses, for example laboratory samples, stars and sellar atmospheres.

The picture of discrete energy levels for individual molecules changes when there is a large number of molecules close together in a gas, liquid or solid. For a solid the discrete energy levels begin to overlap into bands of allowed energy with gaps of forbidden energies in between, in accordance with the theory of quantum mechanics. The local kinetic energy of individual atoms or molecules can cause the emission or absorption of a photon with a little more or a little less energy, shifting its frequency a bit, even if the absorption or emission involves a discrete energy level[11]. The result is that, depending on the temperature and density of the gas or solid, the discrete spectrum may be smeared out and a *continuous spectrum* of waves being emitted. Free electrons, for example those in a metal sample or ionized gas, may also lose energy by emission of electromagnetic waves in a continuous spectrum. It is this continuous spectrum that constitutes blackbody radiation.

As we have seen, temperature is proportional to average kinetic energy so we would expect that not all the radiating molecules of a hot object have the same kinetic energy. Fewer will have extremely high kinetic energy and few will have extremely low energies. Emitted photons from these molecules will have a similar distribution with few extremely long wavelengths and few extremely short wavelengths. Max Planck first derived the correct equation for the *irradiance*, $I(T, \lambda)$, of a blackbody, which is the energy per second per square meter given off at a fixed temperature and with a given wavelength. The irradiance of an object at temperature T for a particular wavelength, λ is

$$I(T, \lambda) = \frac{2\pi hc^2}{\lambda^5\{e^{hc/k_B\lambda T} - 1\}}, \tag{2.13}$$

where $k_B = 1.38 \times 10^{-23}$ J K^{-1} is Boltzmann's constant and irradiance is in units of W m^{-2} per wavelength. Equation (2.13) is plotted in figure 2.3 for the same object at four different temperatures. Integrating equation (2.13) over wavelength from zero to infinity gives the Stephan–Boltzmann law given in equation (2.12), which is the total energy radiated (the power) at a certain temperature.

---

[11] As we will see in chapter 9, this spectral broadening is responsible for important infrared absorption bands in the atmosphere.

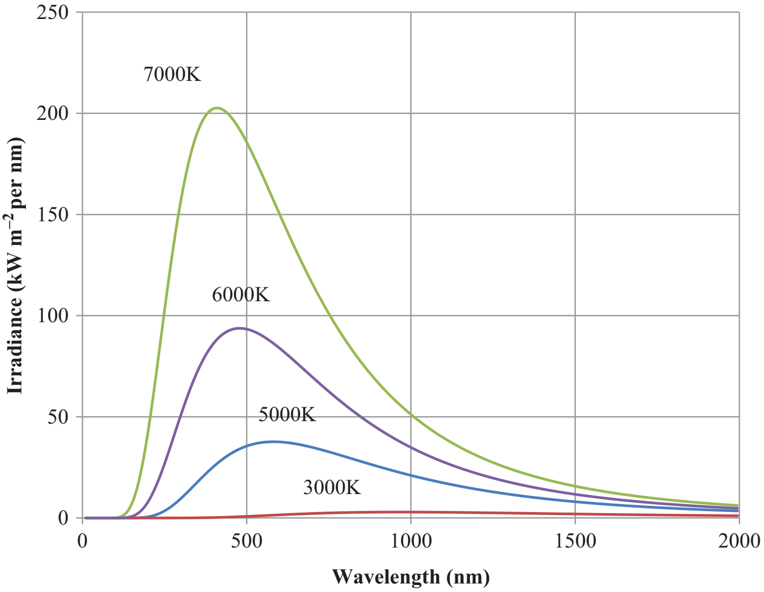


Figure 2.3. Blackbody curves for four identical objects at different temperatures. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

From figure 2.3 we also see that the peak of the spectrum occurs at different wavelengths for objects with different temperatures. Setting the derivative of equation (2.13) equal to zero and solving the resultant transcendental equation numerically for λ gives the peak wavelength as a function of temperature, known as Wien's law:

$$\lambda_{\max} = \frac{2.898 \times 10^{-3}}{T}\ \mathrm{Km}. \tag{2.14}$$

The peak wavelength of the blackbody spectrum can be used to determine the temperature of an object too hot or too remote to be measured with a thermometer. The Sun's surface temperature is 5700 K and the blackbody peak has a wavelength near the color green; most of the emitted energy falls in the visible part of the spectrum. Tungsten light bulb filaments are around 3000 K and most of their energy is given off in the infrared (IR) part of the spectrum. Only a portion of the spectrum is visible, which is why incandescent lamps are only 3% efficient in turning electrical energy into light; most of the energy is given off as heat. Cooler objects also give off electromagnetic waves but not in a part of the spectrum we can see with our eyes. This blackbody spectrum can be detected with cameras, however, which is the basis of some kinds of night vision technology. Living creatures are warmer than their surroundings and so emit a different blackbody spectrum, allowing them to be detected against a cooler background, even when there are very few visible photons. If the planets and their moons were only absorbers of solar energy their surface temperatures would continue to rise. They achieve thermodynamic equilibrium by emitting blackbody radiation, the majority of which is in the IR part of the spectrum. We will look at the energy balance of planet surfaces in chapter 9. Thermal radiation can also be used to passively cool an object if it can radiate into space. If the surface of the object reflects visible radiation so that it does not absorb heat and also emits in the infrared range it will lose energy if exposed to the sky since space is cooler. The effect has been known to cool objects below ambient air temperature on clear nights but has recently been shown to work during the daytime as long as the atmosphere is clear enough to not block too much infrared (Zhai *et al* 2017)

Evaporative cooling, the fourth means of heat transfer, involves the loss of atoms or molecules with higher kinetic energy from the bulk material. This type of process, like diffusion and other mass transfer, is important in energy transfers between open systems. An open container of liquid is more likely to lose the higher speed molecules from its surface, lowering the average kinetic energy and since $\frac{3}{2}k_{\mathrm{B}}T = \langle\frac{1}{2}mv^2\rangle$, the temperature will also be lower. An object can cool by this process, even if there is not an initial temperature difference[12].

An approximate equation for evaporation from the skin is

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = l(P_{\mathrm{s}} - P_{\mathrm{o}}), \tag{2.15}$$

where l is the evaporative heat transfer coefficient with units of W Pa^{-1}, P_{s} is the water vapor pressure adjacent to the skin and P_{o} is the water vapor pressure in the surrounding air. *Vapor pressure*, which is temperature dependent, is the pressure at which the liquid (or solid) and gas phase of a substance come into equilibrium. At this pressure and temperature equal amounts of the substance are evaporating and condensing. The coefficient, l, depends on the air velocity and in general is determined empirically. The water vapor pressure close to the skin's surface, P_{s}, will depend on the temperature of the person's skin and the rate at which the person is sweating, and the vapor pressure of the surrounding air depends also on the temperature and on the humidity. On a hot humid day when P_{o} is high the rate of heat loss due to sweating will be lower, which is why humid climates are less comfortable than arid ones at the same temperature.

2.4 Walls and windows

Approximately 25% of the energy in commercial buildings and 55% of the energy used in homes in the US is used for heating and cooling. Insulation of buildings represents a narrow but significant opportunity to apply basic concepts of heat transfer. For the purpose of calculating heat loss through a composite wall of a building it is convenient to combine conduction and convection into the single empirical equation

$$\mathrm{d}Q/\mathrm{d}t = UA\Delta T, \tag{2.16}$$

---

[12] This does not go against the first or second law of thermodynamics; as we will see in the next chapter the change occurs at the expense of an increase in entropy.

where the constant U, the total heat transfer coefficient per unit thickness, accounts for the possibility of convection on either side of the barrier and conductive energy transfer through the barrier. While thermal conductivity, k, in W $\mathrm{K^{-1}m^{-1}}$ (shown in table 2.2) is an intrinsic property of the material, independent of the size of the sample and surrounding conditions, the U-factor, measured in W $\mathrm{K^{-1}m^{-2}}$ is *not* an intrinsic property of the material and is empirically derived under standardized conditions. An approximate value for U may be estimated using

$$\frac{1}{U} = \frac{1}{h_\mathrm{C}} + \frac{1}{h_\mathrm{H}} + \frac{L}{k} + H, \tag{2.17}$$

where h_C is the convection coefficient on the cool side of the barrier, h_H is the convection coefficient on the hot side, k is the thermal conductivity for the material separating the two substances, L is the thickness of the barrier and H is an empirical constant that depends on the shape and type of surface of the barrier.

In the US the building industry has standardized the reciprocal of the U-factor as thermal resistivity or R-value given in units of $\mathrm{ft^2}$ °F h/BTU inch. Here the British Thermal Unit (BTU) is a measure of energy equal to 1055 J and the values are per inch of a particular insulating material. These are empirical values that can change with conditions, for example moisture has the tendency to lower the R-value of brick that has absorbed rain water. Table 2.3 gives a list of typical building material conductivities and their R values per inch.

The insulating properties of the composite wall of a building can be calculated by adding the R-values of the components. For example a wall composed of a half an inch of drywall, three and a half inches of fiberglass batting, a one inch thick wooden panel and three inches of brick would have an R-value of $0.50 \times 0.90 + 3.50 \times 2.00 + 3.00 \times 0.20 = 8.06$ $\mathrm{ft^2}$ °F h/BTU. Air layers on the inside and outside add 0.70 and 0.20 to this value, respectively. Converting this to the U-factor (1 $\mathrm{ft^2}$ °F h/BTU = 0.176 W $\mathrm{Km^{-2}}$) gives $U = 0.635\ \mathrm{K^{-1}m^{-2}}$. A square meter of wall composed of this material will then allow a heat transfer of $\frac{\mathrm{d}Q}{\mathrm{d}t} = UA\Delta T = 0.635 \times 1.0\ \ \mathrm{m^2}$ (20.0 °C) = 12.7 W for a temperature difference of 20 °C between the inside and outside temperatures.

Buildings cannot be sealed up completely as the build-up of moisture from condensation and carbon dioxide would eventually make the structure uninhabitable. A comfortable exchange rate of air for a room is about 1.5 exchanges per hour which is around eight liters of fresh air per person per second. In the US recommended ventilation rates for commercial kitchens are 30 exchanges per hour, for restaurants 10/h, and for classrooms and offices, 4/h. Ventilation requirements result in energy losses of around 15% for a typical building. In a typical building the roof accounts for 25% of the energy loss, windows about 10%, walls 35% and floors 15%. These figures will vary depending on the construction and can be somewhat ameliorated using heat exchange systems (called Heat Recovery Ventilation—HRV) which use outgoing warm air to pre-warm incoming cool air.

Even with ventilation losses it is possible to insulate a building well enough that internal sources are sufficient to supply the entire heating needs of a structure.

Table 2.3. Approximate R-values for various common substances. See Monteith and Unsworth (2008).

Material	R-value in $\mathrm{ft}^2\ {}^\circ\mathrm{F\ h/BTU\ inch}$
Vacuum	30
Thinsulate (clothing)	5.75
Air (no convection)	5.75
Soil	0.1
Water	0.24
Stainless steel	0.01
Aluminum	0.0006
Copper	0.00036
Polyurethane foam	3.6
Wood panel	2.5
Bale of straw	1.45
Snow	1
Fiberglass, loose	2.2
Fiberglass batting	2
Polyisocyanurate panel with foil	6.8
Cellulose, loose	3
Cardboard	3
Brick	0.2
Poured concrete	0.1
Drywall	0.9
Windows	
Single pane	1
Double pane	2
Triple pane	3
Double pane with coating	4

Humans generate about 100 W if seated, up to 140 W if active, lighting provides 20 W to 40 W per square meter, computers about 150 W, televisions 100 W and photocopiers up to 800 W. With these sources, good insulation and passive solar energy it is possible to make buildings which do not need central heating units, even in cold climates. Painting the roof of commercial buildings with an IR reflecting coating has been shown to lower air conditioning costs and reduce regional heat island effects by 1.2–1.6 °C (Zhang *et al* (2016), and references therein).

Windows present a special problem in terms of preventing heat loss from a building because the material must also be transparent in addition to having favorable thermal transfer characteristics. Before the early 1970s cooling and heating losses through windows in the US consumed around 5% of the total energy used (Rosenfeld *et al* 2000). Improved window design has lowered this loss to 3%. This has been accomplished by changing the chemical makeup of the glass itself, making windows with multiple glazing (several layers) with non-conducting gases between them and glass coatings that allow visible light to pass through but block IR wavelengths.

The thin layer of air adhering to a window pane accounts for 20% of its insulating capacity but is greatly reduced if there is forced convection in the form of wind as noted above. Multiple glazing and storm windows stabilize these layers so that they maintain their insulating capacities even in windy conditions. Gases with better insulating capacity than air can be chosen to fill the space between the glass and other transparent materials such as *aerogel* are under investigation.

The inside of a building can be thought of as a blackbody source in the IR with most of the energy having wavelengths greater than 1000 nm. This accounts for two thirds of the energy loss through a double glazed window. Conventional glass transmits visible light but absorbs IR and then re-emits this energy to both the inside and outside. The first window coatings on the market blocked nearly all wavelengths (and consequently were very hard to see out of) but newer coatings have been developed that selectively allow visible to pass through but reflect IR. These coatings have reduced the effective U-factor of windows by 30%. Window coatings for southern climates are now adjusted to block all IR but in northern climates coatings are often made to admit near IR wavelengths (1000 nm to 3000 nm) while blocking other IR. This makes it possible to take advantage of solar heating during the winter in northern climates when the Sun is at a much lower angle in the sky.

Current research in window technology involves making systems which dynamically adjust to ambient conditions. Transition eyeglasses change tint on exposure to sunlight via the *photochromatic* effect whereby UV wavelengths cause a temporary chemical change in silver compounds embedded in the glass. This technology may eventually become sufficiently inexpensive to use in windows. A similar effect called *electrochromism* uses chemical compounds in glass which are sensitive to an applied voltage to change the amount of light passing through the window. This process allows the window to be adjusted from highly reflective to all wavelengths to selectively transparent to various ranges of wavelengths.

2.5 Lighting

The numbers shown in table 2.1 for the efficiencies for different kinds of lighting are determined by the total energy output at all frequencies divided by the energy needed to operate the lamp. However a significant portion of the light given off by most light sources is not in the visible portion of the spectrum and it might be more useful to know how much usable light is provided for a given input of energy. The *lumen* is a subjective measure of light intensity as seen by the human eye, adjusted for frequency since the human eye does not have the same sensitivity at all wavelengths. *Luminous efficacy* is the ratio of lumens emitted per watt used and most light bulb packaging now includes this figure instead of or in addition to energy efficiency. Approximately 20% of the electricity generated in the US is used for lighting making the topic of lighting efficacy a significant one. The efficacy for a selection of commercial lighting is shown in table 2.4 and these values should be compared to the efficiencies for lighting given in table 2.1.

The light emitted by incandescent, halogen and arc lamps is blackbody radiation due to the high temperature of the material inside the bulb. As mentioned above,

Table 2.4. Lighting efficacies for various commercial lighting.

Type of light	Efficacy (lm W^{-1})
Candle	0.3
Incandescent	5–12
Halogen	16–24
Fluorescent, compact	45–60
Fluorescent, tube	50–100
Xenon arc	30–50
Sodium discharge	150–200
Sulfur discharge[a]	100
LED	26–100[b]
LED prototypes	300

[a] In a sulfur discharge lamp the sulfur is excited by microwaves rather than directly by electrical currents.

[b] Efficacy for LEDs depends on color, with orange, green and cyan lights having slightly higher efficacies than blue or red.

incandescent lamps have tungsten filaments at around 3000 °C so that only part of the energy emitted is in the visible spectrum, the rest is in the IR part of the spectrum and so is perceived as heat. Higher temperatures would produce more visible light but the tungsten tends to vaporize if temperatures are too high. In normal incandescent bulbs the filament is in a vacuum but slightly higher temperatures can be achieved without melting the tungsten filament by putting a small amount of a halogen gas such as bromine or iodine in the bulb. Tungsten from the filament will still vaporize and be deposited on the inside of the glass bulb, but the halogen atoms act to return the tungsten to the filament, thus preserving the filament from disintegrating. In arc lighting there is not a filament but an electric current flows through a gas between two electrodes. The gas can be heated to a much higher temperature than any solid, shifting the peak of the blackbody spectrum towards the visible.

Fluorescent lamps and other gas discharge tubes operate at much lower temperatures and use discrete spectrum light instead of blackbody radiation. When an electric current passes through a gas such as neon or mercury it excites the electrons in such a way that they move to discrete higher energy levels. As the electrons return back to the lower energy levels photons of a characteristic wavelength of light are emitted ($f = E/h$ where E is the energy difference between the two electron energy levels and h is Planck's constant). The discrete spectral lines of neon are predominantly in the orange/red part of the spectrum, giving a neon sign its distinguishing reddish orange color. The sodium spectrum has two predominant yellow spectral lines and mercury has several lines in the blue part of the visible spectrum which results in the distinctive colors of street lights using these gases.

A gaseous mixture of mercury, argon and neon excited by a current will cause the mercury atoms to give off a discrete spectrum in the UV range. Commercial and home fluorescent lighting is made by coating the inside of a mercury vapor lamp

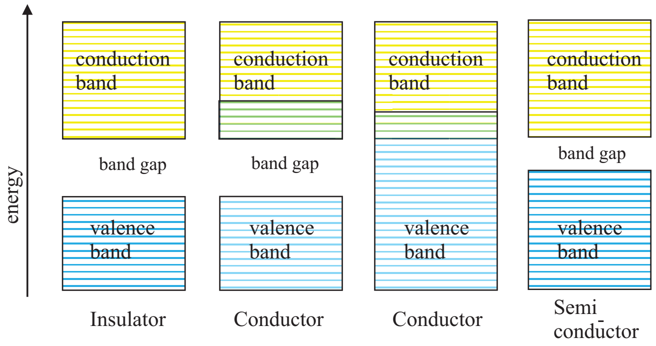


Figure 2.4. Blue shading represents energy levels full of electrons (the valence band), yellow shading represents energy levels which are empty and available for electrons to move to (the conduction band) and green shading shows an overlap of valence and conduciton bands. Energies in the band gap are forbidden to electrons as the result of quantum mechanics. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

with a material that will absorb UV and re-emit visible light. The UV light is absorbed causing electrons to go to higher energy levels but the electrons return to the lower level in a two-step process, one of which produces visible photons. The process of absorption at high frequencies (low wavelengths) and re-emission at lower frequencies (higher wavelengths) is called *fluorescence*. Tubes without coatings will cause clothing or other material to fluoresce and the effect is sometimes used in bars or night clubs and in theater productions. The visible spectrum given off by a coated lamp can be adjusted by combining compounds with different visible spectra giving rise to ‘cool white’, ‘natural white’ or other colors. Fluorescent lamps also have a blackbody spectrum (as do all objects above zero kelvin) but it is the discrete spectrum which is used to provide visible light and the tube can operate at a much lower temperature as a result. *Phosphorescence* occurs when the visible photons are delayed, sometimes more than an hour, before being emitted. This makes glow-in-the-dark signs, toys and watch dials possible[13].

Solid state devices called light emitting diodes (LEDs) are another mechanism for turning electrical current into visible light. As was noted previously, in a solid the energy levels available to an electron become smeared out into bands of energy levels. The exact structure of these bands and the number of available electrons in them depends on the chemical properties of the atoms making up the material and can be calculated using quantum mechanics. The four basic band/gap possibilities are shown in figure 2.4. A material is an insulator if electrons completely fill a lower energy band, called the valence band, and there is a large gap so that no electrons can reach the next energy band. In this case the electrons are trapped in an energy band with no possibility of being able to move. If the next highest band of energies, called the conduction band, is partially filled, a large number of electrons can change

---

[13] Early glow in the dark watch dials used radioactive materials to cause a glow. This technique is no longer used due to the danger of accidental radiation exposure.

energies very easily when a voltage is applied and the material is a conductor. A conductor also results if a filled lower band overlaps with an empty band so that there are plenty of free energy levels available to the electrons and they can move around in the solid. In the case where the valence band is filled with electrons but is close enough to the empty conduction band that at least a few electrons can jump across (for example due to thermal kinetic energy and Heisenburg's uncertainty principle which introduces a probabilistic distribution of energies for the electrons.) the material is an intrinsic semiconductor[14].

The picture shown in figure 2.4 is further complicated by the fact that additional energy levels can be produced in the gap for a particular substance by adding trace elements to the material, a process called *doping*. If the doping leads to extra filled energy levels in the gap just below the bottom of the conduction band the material is called an n-type semiconductor (n for negative charge carriers). In n-type semiconductors thermal energy is enough to promote the electrons in the new energy levels into the conduction band where they can act as charge carriers in a current flow. If the doping process ends up creating unfilled energy levels just above the top of the valence band, electrons will move to these energy levels from the valence band leaving behind vacancies. These holes in the valence band act as positive charge carriers and the material is said to be a p-type semiconductor.

This inherently quantum mechanical behavior of energy bands in doped semiconductors has been exploited in many modern electronic devices such as transistors, diodes, LEDs and solar cells. Basic solar cells and LEDs involve joining a piece of n-type silicon with a piece of p-type silicon, a combination called a *p–n junction.* Electron flow tends to occur only in one direction because the energy levels available to electrons on one side of the junction are lower than on the other. A simple p–n junction acts as a diode in an electric circuit, allowing current flow through it in one direction but blocking current trying to flow in the other direction. Figure 2.5 shows

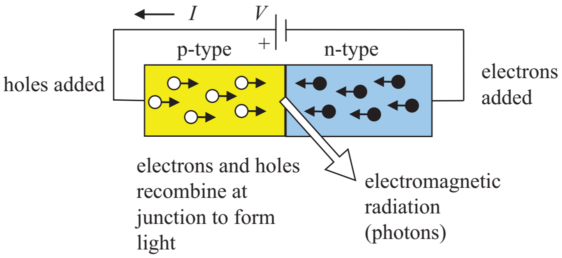


Figure 2.5. An applied voltage will cause electrons (traveling in the opposite direction to the conventional current, I) to cross the n-type semiconductor to the junction where they will fall into lower energy holes on the p-type side, giving off photons. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

---

[14] The gap for most semiconductors is actually quite a bit larger (>1 eV) than the average electron thermal energy at room temperature (on the order of 0.025 eV), but because there are so many electrons the distribution of thermal energies is broad enough that a significant number can span the gap. Only about one out of 5×10^{12} electrons is promoted to the conduction band in silicon at room temperature but this is sufficient for significant current flow.

a schematic diagram for a p–n junction acting as a LED where electrons are added to the n-side of the junction via a current flow and holes are added (electrons removed) from the p-side. When the holes and electrons recombine at the junction the energy of recombination is emitted as electromagnetic radiation. This process can also be engineered to operate in reverse, in which case incoming photons cause new electrons and holes to form at the junction and a current will flow through an external circuit. This is the basis of photovoltaic (PV) solar cells, which will be discussed in more detail in chapter 6.

The width of the band gap determines the energy of the photons given off and the manipulation of this gap has been a central focus of research for the past few decades. For a band gap of energy E the frequency of photon given off is determined by $E = hf = hc/\lambda$ where E is the energy of the band gap, f is the frequency of the emitted photon in hertz, c is the speed of light and h is Planck's constant. The first LEDs made had band gaps in the range of 13 eV which gives rise to IR light at around 950 nm. The physical structure of the crystals involved can be adjusted to make solid state *lasers* which are used to read and write CDs. The ability to create DVDs, which hold more information compared to a CD, had to await the development of blue LED lasers which have a higher resolution due to smaller wavelengths which required smaller band gaps. White light from an LED can be produced in much the same way as for fluorescent bulbs by producing UV light which is re-emitted in the visible range by various coatings. Three colors of LEDs can also be combined in different amounts to achieve the other colors and hues needed for television and computer monitors. In theory these monitors should use less energy than an equivalent sized TV using older technology, although often other factors, such as the amount of electricity used in standby mode and larger screen sizes, can actually make newer technology *less* energy efficient.

LEDs built with organic compounds (OLEDs) are currently used in digital displays and laboratory prototype versions of white OLEDs with very high efficacies exist. Greater efficacies, longer lifetimes, resistance to shock, small size and the fact that they do not contain mercury make LEDs an attractive light source[15].

2.6 Energy transfer using electricity

Current flow is not technically a heat transfer mechanism but it is an important means of moving energy from one place to another. As mentioned previously, traditional electric generating coal plants are about 32% efficient in the US which means every 1 kWh of electrical energy consumed requires burning approximately 3 kWh of primary energy in the form of coal. These losses are the result of limitations imposed by the second law of thermodynamics which will be discussed futher in the next chapter. An additional 7% is lost in transmission from the generating plant to where the energy is used.

---

[15] The manufacturing of LEDs and solar cells involves toxic chemicals but in the finished product the dopant materials are chemically bound into non-toxic compounds and are not easily separated from the bulk material.

The amount of charge, measured in coulombs, passing a point in a wire per second defines the *current*, I, measured in amperes, A, which is a coulomb per second. Negatively charged electrons are the actual charge carriers in most cases but electrons were not proven to exist until long after many engineering applications, with the result that the conventional current shown in circuit diagrams flows from the positive terminal of a battery to the negative which is opposite to the direction the electrons actually travel[16].

The energy per charge of the charge carriers is measured in volts where a *volt* is defined to be equal to a joule per coulomb. An applied voltage gives the electrons potential energy, much like lifting a mass in a gravitational field gives a mass gravitational potential energy or pressure in a pipe makes water flow if the tap is turned on. The amount of current flow through a resistor depends on the voltage and the resistance and is given by Ohm's law: $V = IR$ where R is the *resistance* measured in ohms.

The operation of many electrical devices, such as lights, toasters, appliances, etc, can be approximated as a series of resistances, measured in ohms. Resistance occurs when electrons interact with the atoms that make up the material through which they are flowing[17]. The kinetic energy lost by the electrons becomes random kinetic (or thermal) energy that causes the light bulb to light, the toaster to heat up, etc. The same number of electrons per second (current) flows out of a resistance as flows into it but the charge carriers have less energy when they leave than when they started.

Because of the interaction of electrons with the atoms in a conducting wire the wire gains random thermal energy and its temperature will rise. This is the major source of the 7% energy lost in transmitting electricity. The power lost by a resistor due to heat flow out of it is proportional to current squared,

$$P = RI^2. \tag{2.18}$$

There are other losses in electrical transmission cables such as electromagnetic waves given off by alternating current but resistance loss is the largest.

Power plants using coal, natural gas, nuclear energy, geothermal sources or solar thermal energy all operate in basically the same manner by making steam. This steam is used to turn a turbine which turns a generator. Hydroelectric plants use flowing water to turn a turbine which turns a generator and modern windmills use wind to turn the generator directly. In all these cases electricity is produced by the generator as the result of Faraday's law which says a changing magnetic flux will cause an electrical potential,

$$\varepsilon = -\frac{\mathrm{d}}{\mathrm{d}t}\int \vec{B} \cdot \mathrm{d}\vec{A}, \tag{2.19}$$

where ε is the electromotive force (emf), measured in volts and A is some area where there is a changing magnetic field. The term emf is synonymous with voltage and for

---

[16] An electron carries a charge of 1.6×10^{-19} C.

[17] The interaction is quantum mechanical in nature but can be approximated by the Drude model which considers electrons as small balls ballistically colliding with stationary atoms in the wire.

calculation purposes emf and voltage, V, are interchangeable. The dot product measures how much magnetic field penetrates the area, which is called the magnetic flux[18]. Most modern power generators involve large stationary electromagnets with large coils of wire mounted on a shaft which is turned by the turbine so that it is the angle between the area enclosed by the coil which changes orientation in the static magnetic field. A rotation rate of 60 revolutions per second creates the 60 Hz alternating current used in the US (50 Hz in Europe). Credit card readers, magnetic tape players, the read head of a magnetic computer disk drive, electric guitar pickups, traffic sensors embedded in the highway and transformers are other devices that depend on Faraday's law to operate. In all these cases a changing magnetic field causes a voltage which then causes a current flow via Ohm's law.

Faraday's law and the use of alternating current allow a great reduction in the resistive power loss given by equation (2.18). In a transformer a large block of iron or other magentic material (which intensifies magnetic field by lining up the atomic dipoles in the material in regions called *domains*) is wound with two separate coils of wire, each with a different number of turns. Alternating current enters the transformer through one set of loops called the *primary* and this creates a changing magnetic field in the second coil called the *secondary*. Via Faraday's law, the changing magnetic field in the secondary coil causes a current to flow in it. The voltage on each side of the transformer is proportional to the number of turns of wire on each side:

$$\frac{V_S}{V_P} = \frac{N_S}{N_P}, \tag{2.20}$$

where V_P and V_S are the voltages across the primary and secondary and N_P and N_S are the number of turns of wire in the primary and secondary, respectively. Energy must be conserved so the power, $P = VI$, in the two coils must be equal or

$$P_P = V_P I_P = V_S I_S = P_S. \tag{2.21}$$

What these two equations show is that, by using different numbers of turns in the primary and secondary coils, voltage can be stepped up while current is stepped down with the same power transmitted. Transformer stations routinely change the electric current from a generator at the generation plant to very high voltage, low current flow so that it can be transmitted with less resistive power loss (2.18). At the other end of the process, close to where the current will be used, another transformer steps the current up and voltage down. Some energy is lost as very low frequency electromagnetic waves by the alternating fields in the transformer and in the transmission wires, but this loss is more than offset by the savings in transmitting energy with less resistive power loss. For example, a 100 W light bulb operating on a 110 V power supply (the standard in the US) will draw a 0.9 A current ($P = VI$) and use 100 W × 7 days × 24 h × 3600 s = 6.0×10^7 J in a week of continuous operation. Transmitted at 110 V and 0.9 A the power loss will be 1.6 W if the transmission line has a total

---

[18] Recall that the area vector is defined to be perpendicular to the area and the dot product is given by $\vec{B} \cdot \vec{A} = BA\cos\theta$ where θ is the angle between the two vectors.

resistance of 2.0 Ω.[19] Using a transformer to increase the transmission voltage to 10 000 V will reduce the current to 0.01 A and the power loss to 2.0×10^{-4} W. The energy savings over one week of operation is 121.0 J versus 9.7×10^{5} J.

Projects

1. Look up the history of the 'water-fueled car' on the Internet (the earliest reports of this scam go back to 1935). List some of the various claims made by these inventors and show how the first law of thermodynamics would be broken if the claims were true. Combine your analysis and data into a report.
2. Find the efficiencies of conversion processes not listed in table 2.1 and compare them to the figures in the table. Have there been improvements in the table figures? Write an analysis of what distinguishes highly efficient conversions with processes that are lower in efficiency.
3. Design a simple building using materials listed in table 2.3 and calculate the heat loss from the building for a typical winter day where you live. Compare the various ways of heat loss; which is most significant? Would the appliances in the building give off enough energy to keep the building warm? How much additional insulation would be required to supply all of the heat from people and devices inside the building?
4. Go to the MIT web site http://designadvisor.mit.edu/design/ and experiment with their online software for designing a building. Design a building and report on your conclusions.
5. Find an estimate for the number of windows in your country, city or state. Assume a reasonable size for these windows and calculate the typical heat loss through them in winter under reasonable estimates as to their size and the outside temperature. How would this loss change if all of these windows were triple-pane instead of single sheets of glass?
6. Write a summary of the Leadership in Energy and Environmental Design (LEED) program. What are the standards? How have they been applied?
7. Look up the cost, typical lifetime and energy use for an incandescent, fluorescent and LED lamp which produce the same lumens of light. Make a plot comparing the cost over time of these devices. At what time does the greater efficiency of an LED lamp win over the smaller initial cost of the incandescent? A metric tonne of coal has about 8200 kWh of energy. If the coal plant is 35% efficient in converting the primary energy into electricity, how much coal is used to run each bulb over a year's time? Estimate the number of light bulbs being used in your city. How much coal is saved over a year if all of the bulbs are changed from incandescent to LED?
8. Explore the efficiencies of various household appliances over time (refrigerators, television sets, radios, cell phones, etc). It has been claimed that

---

[19] Typical power lines might have a resistance of 0.4 Ω km^{-1}.

efficiency gains from better refrigerators and air conditioning has been offset by the rise in the use of personal electronics. See if you can either verify or dispute this claim with reliable data.

9. Find the peak wavelength of light coming from the Sun (surface temperature 5780 K). What color is this? Discuss color choices of objects where their visibility is important such as tennis balls, safety vests worn by road construction workers and emergency vehicles. Also look up the color sensitivity of the eye. How is this related?
10. Investigate the conflicting claims that (i) fluorescent lights will increase mercury levels in the environment when they are disposed of and (2) the energy savings offered by fluorescent lights will reduce mercury levels in the atmosphere because less coal will be burned, which will mean fewer coal associated mercury emissions. Write a report that explains which claim is right.
11. Estimate the energy savings for transmitting all residential electricity in your state or city at 100 000 V versus 110 V over a year's time.
12. There are many websites available which give advice for reducing your personal or home energy use. Compare advice from several sites. Do they agree on priorities? Analyze their advice based on what you have learned in this chapter.

References

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Jacobsen M Z, Delucchi M A, Bazouin G, Bauer Z A F, Heavey C C, Fisher E, Morris S B, Piekutowski D J Y, Vencill T A and Yeskoo T W 2015 100% clean and renewable wind, water, and sunlight (WWS) all-sector energy roadmaps for the 50 United States *Energy Env. Sci.* **8** 2093

Kromdijk J, Głowacka K, Leonelli L, Gabilly S T, Iwai M, Niyogi K K and Long S P 2016 Improving photosynthesis and crop productivity by accelerating recovery from photoprotection *Science* **354** 857

Monteith J L and Unsworth M H 2008 *Principles of Environmental Physics* 3rd edn (New York: Elsevier)

Rosenfeld A H, Kaarsberg T M and Romm J 2000 Technologies to reduce carbon dioxide emissions in the next decade *Phys. Today* **53** 29

Smil V 1999 *Energies, An Illustrated Guide to the Biosphere and Civilization* (Cambridge, MA: MIT Press)

Zhang J, Zhang J L and Ban-Weiss G 2016 Revisiting the climate impacts of cool roofs around the globe using an Earth system model *Env. Res. Lett.* **11** 084014

Zhai Y, Yaoguang M, Sabrina N D, Dongliang Z, Runnan L, Gang T, Ronggui Y and Xiaobo Y 2017 Scalable-manufactured randomized glass-polymer hybrid metamaterial for daytime radiative cooling *Science* Feb. 9

Zhu X, Long S and Ort D 2008 What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Curr. Opin. Biotechnol.* **19** 153

IOP Concise Physics

Physics and the Environment

Kyle Forinash III

Chapter 3

Entropy and the second law of thermodynamics

The efficiency of an energy conversion process was defined by equation (2.3) in the last chapter as $\eta = \frac{\text{energy benefit}}{\text{energy cost}}$, where energy benefit represents the useful work done in some process (often a cyclic process) and energy cost is the energy input necessary to achieve the benefit. Efficiency can never be larger than 1 (100%) because of the first law of thermodynamics which says energy is conserved. Table 2.1 gave the efficiencies for a number of energy conversion process and it was noted that some processes are inherently more efficient than others. The topic of this chapter is the *second law of thermodynamics*, which explains why efficiencies cannot reach 100%, even in theory. For processes involving the conversion of heat into mechanical or electrical energy (heat engines) the second law is particularly restrictive as to the maximum possible efficiency.

There are several equivalent ways of stating the second law, each of which applies to closed thermodynamics systems, where all energy flows into and out of the system are accounted for and cyclic systems where the internal energy returns to the same value at the end of the cycle. It is possible to start with any of the three following statements of the second law and prove the other two must be true.

1. Entropy remains constant or increases in a closed system.
2. In a closed system heat flows from hot to cold; it takes input energy to make it flow the other way.
3. The Carnot cycle is the most efficient cycle possible for a reversible heat engine operating in a cyclic process:

$$\eta = \left(1 - \frac{T_\mathrm{C}}{T_\mathrm{H}}\right) \quad \text{(Carnot cycle)}. \tag{3.1}$$

doi:10.1088/978-1-6817-4493-3ch3

3.1 Entropy remains constant or increases in a closed system

We start by defining *entropy*. The 16 possible results of tossing four fair coins, where there is a 50/50 chance of heads or tails, are shown in table 3.1. The coins are distinguishable in the sense that we can keep track of which coin is which during each toss.

From the results we can see that some outcomes are more likely than others. Out of 16 possible outcomes, there is only one way to have all heads but six ways to have two heads and two tails. The number of different ways the same event can happen is called the number of microstates or the *multiplicity*, Ω which has no units. For the case of four coins there is only one microstate available for all heads (or all tails) and the multiplicity $\Omega = 1$. The multiplicity of two heads and two tails is $\Omega = 6$ because there are six distinct ways for this to occur.

Entropy is defined to be

$$S = k_{\mathrm{B}} \ln \Omega, \tag{3.2}$$

where k_{B} is Boltzmann's constant ($k_{\mathrm{B}} = 1.38 \times 10^{-23}$ J K^{-1}), Ω is the number of available microstates (the multiplicity) and ln is the natural logarithm. The entropy for having all heads ($S_{\text{all heads}} = k_{\mathrm{B}} \ln 6 = 2.47 \times 10^{-23}$ J K^{-1}) is larger than the entropy of having one head ($S_{\text{one head}} = k_{\mathrm{B}} \ln 1 = 0$ J K^{-1}) so from an entropy standpoint, final states with higher entropy (two heads and tails) are more likely. The first version of the second law is a claim about probabilities; the more probable state (defined to have a higher entropy) is more likely for a closed system. If we continue to toss the coins we will see the highest entropy state more often.

For a small number of coins there will be fluctuations from the most likely outcome—throwing four coins does sometimes result in all heads. But this picture changes drastically if the number of objects increases significantly. For larger numbers of objects that can have two possible states it becomes unmanageable to write down all the possible outcomes and probability theory must be used (Glazer and Wark 2001 is a useful guide to statistics related to thermodynamics). The number of possible outcomes for N objects when there are two states (heads and tails for coins) is given by 2^N and the probability of the number of heads for a given throw is $p_{\mathrm{H}} = N/2$ (the most likely outcome is half heads and half tails). Probability theory

Table 3.1. All possible outcomes of tossing four coins.

Coins	Toss 1	Toss 2	Toss 3	Toss 4	Toss 5	Toss 6	Toss 7	Toss 8	Toss 9	Toss 10	Toss 11	Toss 12	Toss 13	Toss 14	Toss 15	Toss 16
1	H	H	H	H	T	H	H	H	T	T	T	H	T	T	T	T
2	H	H	H	T	H	H	T	T	H	H	T	T	H	T	T	T
3	H	H	T	H	H	T	H	T	T	H	H	T	T	H	T	T
4	H	T	H	H	H	T	T	H	H	T	H	T	T	T	H	T
		Three heads				Two heads						One head				

also tells us that the number of ways of arranging N objects between two possible states (H and T) with N_H objects in state H is given by the binomial distribution,

$$\Omega_{N_\mathrm{H}}^{N} = \frac{N!}{N_\mathrm{H}!(N-N_\mathrm{H})!}, \tag{3.3}$$

where $x! = x(x-1)(x-2)\ldots1$ is the factorial function (for example $4! = (4)(3)(2)(1) = 24$)[1]. For the case of four coins equally divided between two possible states we have $\Omega_2^4 = 6$, a result we obtained above by simple counting. We also obtain $\Omega_0^4 = 1$ for no coins in the H state (where we have used the fact that $0! = 1$).

These ideas can be applied to large numbers of molecules in a gas sample and this forms the basis of the science of statistical mechanics (see for example Schroeder (2000) from which this discussion was adapted). Suppose we wish to look at multiplicities for 40 molecules and the two states available to them are the two halves of a container (side H and side T). We have $\Omega_0^{40} = 1$ or one way to find them all on one side but now we have $\Omega_{20}^{40} = 1.38 \times 10^{11}$ different ways to split the molecules so that half are on each side. The total number of possible outcomes is $2^{40} = 1.10 \times 10^{12}$ so there is a one in 1.10×10^{12} chance of finding all 40 molecules in one state. However there is a $1.38 \times 10^{11}/1.10 \times 10^{12} = 0.125$ or 12.5% chance of finding the molecules equally split between the two states compared to a $1/1.10 \times 10^{12} = 9.09 \times 10^{-13}$ or $9.09 \times 10^{-11}\%$ chance of finding them all on one side.

A plot of the available multiplicities for a system where the outcomes are random will form a normal distribution with the largest multiplicity (also the average multiplicity in this case) in the center surrounded by the other possible multiplicities. In probability theory the *standard deviation*, σ is used to define the spread of a distribution. Approximately 68% of the values will fall within $\pm\sigma$ of the average (or mean) value, around 95% of the values will be within $\pm2\sigma$ of the mean and 99.7% of the values will be within $\pm3\sigma$ of the mean. For the binomial distribution (3.3) $\sigma = \sqrt{Np(1-p)}$ where p is the probability of a particular outcome. For two state systems $p = 0.5$ or 50% assuming the two states are equally likely. The standard deviation for $N = 4$ is $\sigma = 1$ and the height of the distribution is $\Omega_{N/2}^{N} = 6$. For $N = 40$ we have $\sigma = 2.24$ and the height of the distribution is $\Omega_{N/2}^{N} = 1.38 \times 10^{11}$. Notice that for 40 objects the distribution is extremely narrow compared to its height, whereas the distribution for four objects has a standard deviation of 1/6 the height. We expect to occasionally see all four coins come up heads in a four coin toss but it is exceedingly unlikely that all 40 molecules will end up on one side of the container in a system of 40 molecules. For larger numbers of molecules, for example the number of molecules in the room you are in right now, we never have to worry that suddenly all the molecules will wander to one side of the room leaving you without air to breath.

---

[1] Most calculators and computers are limited to calculating factorials of numbers less than 500 because the values become very large. For higher numbers of entities the Stirling approximation and other mathematical tricks must be used to evaluate equation (3.3). See Glazer and Wark (2001) for more details.

The above example of a two state system can be generalized to systems with many possible outcomes but the conclusion is the same. For very large numbers of objects, for example a mole (6.02×10^{23} entities), the system is extremely likely to end up in the state with the greatest entropy. Nothing in the first law of thermodynamics prohibits all of the molecules in a room from randomly wandering into a corner of the room, leaving a vacuum in the rest of the room, however, the second law says that this is so improbable that it will never occur (or possibly occur once in a time larger than the age of the Universe).

3.2 In a closed system heat flows from hot to cold

A simplified model system with more than two states which can be examined using the concepts of probability is an Einstein solid. In this model the energy levels are quantized so that only exact whole number units of energy can be added or subtracted from the system. The binomial distribution (3.3) does not apply but similar laws of probabilities hold. The available energy states are assumed to have equal probability and are equally accessible, a property sometimes referred to as the *fundamental assumption of statistical mechanics*. Let us start with a solid which has only three atoms ($N = 3$) and count the number of microstates, Ω, available for a given quantity of added energy to this solid. If no energy is added there is only one available microstate; all three atoms have zero energy and the multiplicity is one. For one unit of added energy the multiplicity is three because we can give the one unit to one atom and there are three possible atoms which might have it. For two units of energy added there are six microstates. Table 3.2 summarizes the multiplicities and possible distributions of energy.

Probability theory gives the general formula for finding the multiplicity for a given number of oscillators, N, distributed with a given number of energy units, q, as

$$\Omega(N, q) = \frac{(q + N - 1)!}{q!(N - 1)!}. \tag{3.4}$$

Table 3.2. Multiplicity chart for distributing energy units among three atoms in an Einstein solid.

Energy and multiplicity	Atom 1	Atom 2	Atom 3
No energy, $\Omega = 1$	0	0	0
One energy unit, $\Omega = 3$	1	0	0
	0	1	0
	0	0	1
Two energy units, $\Omega = 6$	2	0	0
	0	2	0
	0	0	2
	1	1	0
	1	0	1
	0	1	1

From this formula we see that $\Omega(3,0) = 1$, $\Omega(3,1) = 3$, $\Omega(3,2) = 6$, as in the table above and $\Omega(3,3) = 10$, etc.

Now let us calculate the combined multiplicities for two Einstein solids which have different temperatures (this example comes from Schroeder (2000)). Call the solids A and B and suppose there are N_A atoms in solid A with q_A units of energy and N_B atoms in solid B with q_B units of energy. As a simple example let us suppose each solid has only three atoms or $N_A = N_B = 3$ and there are six energy units to be divided between them: $q_A + q_B = 6$. Initially the solids will be isolated; for now we are just considering the possible ways to split the energy. If A gets no energy, B has all six units; if A gets one energy unit, B gets five, and so forth. Using equation (3.4) above for multiplicity we can fill in table 3.3 for the microstates of each solid. For example in the second row we see solid A has no energy and only one microstate but solid B has all six energy units and 28 different ways (microstates) to spread those six units among the three atoms in solid B.

If the two solids are placed in thermal contact with each other the energy will move around between them with equal probability for each microstate. For the seven cases in table 3.3 we see that the combined number of microstates ($\Omega_{\text{total}} = \Omega_A \times \Omega_B$) for the two solids in contact are $1 \times 28 = 28$, $3 \times 21 = 63$, $6 \times 15 = 90$, $10 \times 10 = 100$, $15 \times 6 = 90$, $21 \times 3 = 63$ and $28 \times 1 = 28$, respectively. The multiplicity for having half the energy in each solid is much larger ($\Omega_{\text{total}} = 100$) than having all the energy in one solid ($\Omega_{\text{total}} = 28$). In other words, assuming the energy can move between the two solids, it is a lot more likely that energy will tend to equalize between them because there are more microstates available for this outcome. Since entropy is defined by the number of microstates (equation (3.2)) the entropy is higher if the energy is distributed equally between the two solids.

From chapter 2 we know that the average kinetic energy of a molecule is proportional to temperature (equation (2.4)) so two Einstein solids with the same average energy per molecule will also have the same temperature. Equalization in energy means equalization in temperature, at least for Einstein solids, and this is the most likely outcome of two solids starting out with unequal amounts of energy. We have thus shown that the first two statements of the second law given at the

Table 3.3. Multiplicity of two Einstein solids of three atoms each sharing six units of energy.

Energy units for q_A	Ω_A	Energy units for q_B	Ω_B	$\Omega_A \times \Omega_B$
0	1	6	28	28
1	3	5	21	63
2	6	4	15	90
3	10	3	10	100
4	15	2	6	90
5	21	1	3	63
6	28	0	1	28

beginning of this chapter are equivalent; heat flow from a hot object to a cool object increases the overall entropy of the combined system and is the most probable outcome, other things being equal. Although we have used a specific example (two Einstein solids in thermal contact) it is true for any two objects (solids, liquids or gas) which can exchange thermal energy with each other but are isolated from their surroundings[2].

As the number of atoms and units of energy increase, an equal division of energy between the two solids becomes *extremely* likely compared to the probability of all, or even most of the energy in only one solid. For example for two solids with 300 atoms each and 100 units of energy split between them the number of microstates with all the energy in one solid is 1.7×10^{96}, a large number, but the number of microstates for an equal distribution of energy is 1.3×10^{122} which is 10^{26} times larger[3]. The mechanisms of heat flow discussed in chapter 2 indicate under what circumstances and how quickly this redistribution of energy will occur, but the laws of probability for large numbers of objects tells us it is *extremely* likely for the system to move towards equally distributed energy, compared to any other configuration of energy sharing.

It is often convenient to think of entropy as *disorder* although there are examples where an association between entropy and disorder does not hold. For the two Einstein solids putting known amounts of energy in each solid constitutes more information about how the energy is distributed than the case where all the energy is distributed between the two. In this case the second law is associated with a trend towards an increase in disorder. In contrast, systems which are not isolated, such as growing organisms, can increase in order because energy flows into the system. The organism lowers its entropy as it grows at the expense of an increase in disorder in the environment. If there is no energy flow into the system the organism will die and decay into disorder.

3.3 The Carnot cycle is the most efficient

As defined in chapter 2, a *heat engine* is a cyclic device that converts heat into mechanical work and returns the system to its original state so that $\Delta U_{\text{cycle}} = 0$. Typical gasoline car engines vary between 10% and 35% efficiency, depending on their size and how they are driven. This means at best, 65 cents of every dollar spent on gasoline for your car is turned into heat that is wasted. Why can we not make a more efficient gasoline engine? New technology can make engines more efficient than those on the market today but the second law places a strict upper bound to the maximum efficiency of any heat engine. The low efficiencies for many of the conversions shown in table 2.1 are the direct result of the second law of thermodynamics.

---

[2] For real substances there are other available energy states, for example rotation and bending motion, and these introduce additional multiplicities. This is in part why different substances have different specific heat capacities. The *equipartition theorem* says the energy will be equalized between microstates.

[3] It should be clear now why it is useful to use the natural logarithm in the definition of entropy. For the first case (all energy in one solid) we have $\ln(1.7 \times 10^{96}) = 221.6$ and for evenly split energy $\ln(1.3 \times 10^{122}) = 281.6$. For multiplicities involving moles of atoms the entropy will be a manageable number after taking natural logarithms and multiplying by Boltzmann's constant.

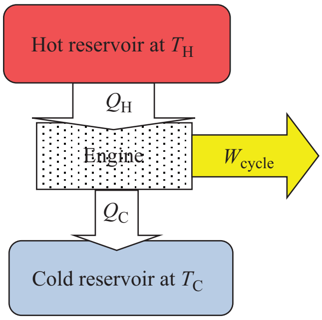


Figure 3.1. Schematic diagram of an ideal heat engine. From Forinash 2010. Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

A schematic diagram for an ideal heat engine is shown in figure 3.1. Energy, Q_H, leaves the hot reservoir, does work, W (the useful work output), and waste energy in the form of heat, Q_C, is expelled to the cold reservoir. For a car engine the hot reservoir is the exploding gasoline or diesel fuel which pushes a piston connected to the drive shaft and the work done is the mechanical rotation of the drive shaft. The cold reservoir is the atmosphere with the radiator and muffler as intermediaries[4]. Because the piston returns to its original location the process is cyclic and the change in internal energy, ΔU is zero for the cycle. Thus for a heat engine, conservation of energy (the first law of thermodynamics, equation (2.1)) gives $Q_H = Q_C + W_{cycle}$.

Efficiency, equation (2.1), for this case becomes $\eta = W_{cycle}/Q_H$. Using $Q_H = Q_C + W_{cycle}$ from the first law, efficiency can be written as

$$\eta = \frac{W_{cycle}}{Q_H} = \frac{Q_H - Q_C}{Q_H} = \left(1 - \frac{Q_C}{Q_H}\right). \tag{3.5}$$

This equation shows efficiency can never be 100%, with the possible exception that either the hot reservoir provides infinite heat or the cold reservoir takes in no heat. An infinite heat transfer from the hot reservoir is clearly impossible, but could the heat expelled to the cold reservoir, Q_C, be zero so that all the input energy is converted to useful work and efficiency is 100%? As shown below, this is exactly what the second law says *cannot* happen because a change in entropy in the wrong direction would be required and as we have seen, decreases in entropy for isolated systems of large numbers of molecules are, for all practical purposes, impossible from a probabilistic standpoint. Intuitively we should suspect this already; we do not expect energy to flow out of the hot reservoir unless it is moving to a cool reservoir.

From the example of two Einstein solids at different temperatures we see that energy will flow from the hot to the cold solid by one of the means of heat transfer until the entropy is at a maximum (the energy is equally distributed). After reaching equilibrium variations in entropy as the result of fluctuations in either heat flow, Q or

---

[4] Radiators and mufflers change the process into a multi-stage heat engine which is beyond the scope of this book.

internal energy, U, must average out to zero[5]. In other words, the change in entropy per change in internal energy equals zero as does the change in entropy per change in heat flow. Therefore at equilibrium we may write $\frac{\partial S_{\text{total}}}{\partial Q_{\text{A}}} = 0 = \frac{\partial S_{\text{total}}}{\partial U_{\text{A}}}$ for solid A where U_{A} is the internal energy and the partial derivatives indicate that we are only interested in how entropy changes with energy or internal energy (the volume or pressure or other conditions will not change because the system is isolated). A similar equation holds for solid B.

The total entropy is the sum of the entropies of the two solids so fluctuations of the total are also zero and $\frac{\partial S_{\text{A}}}{\partial U_{\text{A}}} + \frac{\partial S_{\text{B}}}{\partial U_{\text{B}}} = 0$ at equilibrium. We have assumed the system of two solids is isolated so the energy lost by solid A was gained by solid B and we have $\mathrm{d}U_{\text{A}} = -\mathrm{d}U_{\text{B}}$, which allows us to write $\frac{\partial S_{\text{A}}}{\partial U_{\text{A}}} = \frac{\partial S_{\text{B}}}{\partial U_{\text{A}}}$ at equilibrium. The definitions of temperature and thermal equilibrium tell us the solids must have the same temperature at equilibrium so that $T_{\text{A}} = T_{\text{B}}$ which allows us to formulate a connection between change in entropy with respect to internal energy and temperature: $T_{\text{A}} = (\frac{\partial S_{\text{A}}}{\partial U_{\text{A}}})^{-1}$. The inverse of the entropy change is chosen so that increasing entropy results in lower temperatures if solid A loses heat to solid B. This is in fact a more fundamental definition of temperature than the one relating temperature to average kinetic energy—temperature is defined to be the inverse of the change in entropy as a result of heat flow:

$$T = \left(\frac{\partial U}{\partial S}\right)_{N,V}, \tag{3.6}$$

where the N and V indicate that the number of particles and the volume of the system should be held constant (the system is closed). The temperature, T, quantifies the fact that fluctuations of entropy per change in internal energy are zero at equilibrium.

One very important thing to notice from the above discussion is that *anytime there is an energy flow, Q, there is also a change in entropy* (assuming the number of particles and the volume stay constant). The first law of thermodynamics says that the change in internal energy equals the heat flow plus the work done or $\Delta U = Q + W$. To see how much entropy change there is for a given heat flow when no work is done, we can use the first law and rearrange equation (3.6) as

$$\mathrm{d}S = \frac{\mathrm{d}U}{T} = \frac{Q}{T}. \tag{3.7}$$

Any heat flow requires a change in entropy. This, then explains how evaporative cooling, mentioned in chapter 2, can occur. The increase in entropy of the escaping molecules accounts for the cooling of the liquid left behind.

Using this expression it can be shown that the heat flow to the cold reservoir for our heat engine cannot be zero preventing the construction of any heat engine with 100% efficiency, no matter how clever the engineering. When heat flows out of the

---

[5] This is known as the principle of detailed balance which says that small exchanges of identical molecules and/or energy cannot be detected. It cannot be proved in a strict sense but seems reasonable for large numbers of molecules.

hot reservoir, the entropy of the hot reservoir *decreases* by $dS_H = \frac{Q_H}{T_H}$. There is no entropy change of the engine since a cyclic process returns it to exactly the same state from which it started. So the second law requires that the entropy of the cold reservoir, $dS_C = \frac{Q_C}{T_C}$, has to increase *by at least the same amount as the hot reservoir*. The heat flow Q_C *cannot* be zero in order to maintain constant (or increasing) entropy for the system. The second law requires the total entropy change to be equal or greater than zero or $dS_C - dS_H \geqslant 0$. The best possible case would be an equal exchange of entropy from the hot to cold reservoir, which leads to the maximum efficiency possible.

Substituting the heat flow expressions (3.7) into the inequality $dS_C - dS_H \geqslant 0$ gives $\frac{Q_C}{T_C} \geqslant \frac{Q_H}{T_H}$ or $\frac{Q_C}{Q_H} \geqslant \frac{T_C}{T_H}$. The equal sign applies in the case of *reversible processes* where the steps are infinitesimally small and can be reversed at any stage (think of gradually adding or removing grains of sand to a vertically operating piston). The inequality applies when there is friction or a sudden change in the system (for example a bursting balloon) which is called an *irreversible process*. Using the definition of efficiency in equation (3.5) gives the efficiency of *any* heat engine undergoing a cyclic process as

$$\eta \leqslant \left(1 - \frac{Q_C}{Q_H}\right) = \left(1 - \frac{T_C}{T_H}\right), \tag{3.8}$$

where it is understood that the inequality applies to irreversible processes.

It should be emphasized that these equations apply to *all* heat engines since they were derived with no specification of the engine design involved. The second law expresses a fundamental limitation of nature which cannot be thwarted by technological innovation. Even if we eliminate all friction and all other energy losses, we cannot have any better efficiency than $\eta_{\max} = (1 - \frac{T_C}{T_H})$ for a cyclic heat engine. Nicolas Léonard Sadi Carnot first derived this equation in 1824 and it is referred to as the *Carnot cycle* efficiency. Real engines have less than the theoretical maximum efficiency because they are not Carnot cycles and they lose at least some energy to irreversible processes such as friction. Engines which do not involve heat flows, such as electric motors, will also be limited by the second law of thermodynamics, but in general the constraint is not nearly as severe as is clear from some of the numbers for efficiencies shown in table 2.1 (see the following).

3.4 Heat engines

Figure 3.2 shows the pressure versus volume diagram for the maximum efficiency engine developed by Carnot. In this cyclic process a piston is taken through a series of steps to return to its original volume and pressure. Each step is perfectly reversible so that there is no energy loss. Such a reversible process is defined to be one which is

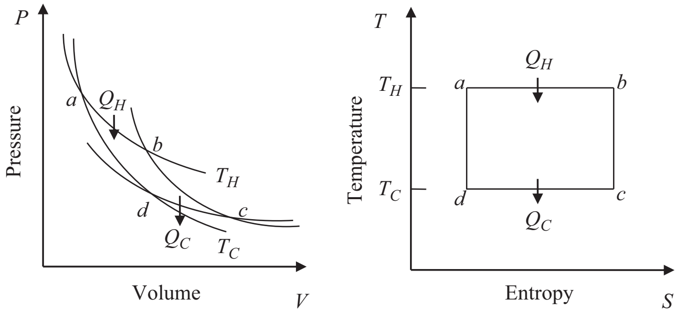


Figure 3.2. Schematic pressure versus volume and temperature versus entropy diagrams for the Carnot cycle. From Forinash (2010). Copyright 2010 Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

quasistatic[6] and frictionless. The net change in internal energy is zero when the piston is returned to its starting point and, as we saw in chapter 2, the area inside the curves will be the work done in the process. The following are the steps in the Carnot cycle:

1. Isothermal[7] heating from $a \to b$. The system absorbs heat, Q_{H}, at constant temperature, T_{H}, by letting the piston expand from P_a, V_a to P_b, V_b while in contact with a large fixed temperature hot reservoir.
2. Adiabatic[8] expansion from $b \to c$. The piston is thermally isolated while it continues to expand from P_b, V_b to P_c, V_c. Adiabatic means there is no heat exchange.
3. Isothermal cooling from $c \to d$. From P_c, V_c to P_d, V_d heat, Q_C, at constant temperature T_C is expelled by letting the piston expand while in contact with a large fixed temperature cool reservoir.
4. Adiabatic contraction from $d \to a$. Thermally isolate the piston and let it continue to contract back to P_a, V_a. No heat is exchanged.

Heat flow (and therefore entropy changes) only occur in the first and third steps. As we saw previously, the entropy flow into the system at the higher temperature has to equal the entropy flow out of the system at the lower temperature so $\mathrm{d}S_{a\to b} = \frac{Q_{\mathrm{H}}}{T_{\mathrm{H}}} = \mathrm{d}S_{c\to d} = \frac{Q_{\mathrm{C}}}{T_{\mathrm{C}}}$. Following the same reasoning leading to equation (3.8) leads us to the same efficiency for the Carnot cycle:

$$\eta = \left(1 - \frac{T_{\mathrm{C}}}{T_{\mathrm{H}}}\right) \quad \text{(Carnot cycle).} \tag{3.9}$$

It is also possible to verify that equation (3.9) comes directly from the definition of efficiency given in equation (2.3), $\eta = W_{\mathrm{cycle}}/Q_{\mathrm{H}}$, by calculating the work done in

---

[6] Quasistatic is another term for a reversible process where the system is in equilibrium at all stages of the process.

[7] Constant temperature.

[8] No heat exchange with the environment. The term *isentropic* is also used for this process meaning that the entropy remains constant. As shown above, a heat flow is equivalent to an entropy change so an isentropic process is also an adiabatic process (one with no heat flow).

each step and combining them to find the total work. Since the work is the area under the curve, each of the four curves in figure 3.2 must be integrated from initial to final points and added (or subtracted for work done on the system) which gives the area inside the curves. The equation of the line on the P–V diagram for the two isothermal processes comes from the ideal gas law:

$$P = nRT/V, \tag{3.10}$$

where n is the number of moles of gas, $R = 8.315$ J mol^{-1}K^{-1} is the ideal gas constant and T is constant. Calculating the work for the process $W = \int_{V_a}^{V_b} P\mathrm{d}V$ gives $W_{a\to b} = nRT \ln \frac{V_b}{V_a}$ with a similar expression for the negative work done in going from $c \to d$.

The equation for the two adiabatic processes is $P = kV^{-\gamma}$ where k is a constant and γ is the ratio of specific heat at constant pressure to specific heat at constant volume (see Schroeder (2000) for a derivation of this equation). Calculating $W = \int_{V_b}^{V_c} P\mathrm{d}V$ in this case results in $W_{b\to c} = \frac{P_bV_b}{1-\gamma}((\frac{V_b}{V_c})^{\gamma-1} - 1)$ with a similar expression for the negative work done in the process $d \to a$. Adding together the work for all four steps and dividing by Q_{H} yields equation (3.9). The calculations for processes other than isothermal or adiabatic are sometimes more complicated but this is in fact how the maximum efficiencies for the real heat engines listed in table 3.4 are calculated. In the case of the Carnot cycle the final result for efficiency can be written in terms of temperature. For other cycles the efficiencies may be expressed as different equations which have temperature, volume or pressure as variables (see the discussion below).

A different and sometimes more convenient expression for the work done by a heat engine undergoing a closed cyclic process can be derived by starting with the first law of thermodynamics, $W = \Delta U - Q$. The total change in internal energy is zero because the system returns to its original state and the heat flow is positive. This allows us to replace heat flow with its entropy equivalent, $\mathrm{d}S = \frac{Q}{T}$ (equation (3.7)) to obtain the following alternative expression for work as the area inside a closed T–S diagram:

$$W_{\mathrm{cycle}} = \oint T\mathrm{d}S. \tag{3.11}$$

Here the circle on the integral sign reminds us that we must add the entropy changes over an entire cycle. For the Carnot cycle the T–S diagram is much simpler than from a P–V diagram as seen in figure 3.2.

The Carnot cycle has the greatest efficiency of any heat engine but practical engines do not use the Carnot cycle for a number of reasons. The requirement that the steps be quasistatic in effect means the process must be done very slowly, which is not suitable for practical applications. Real engines also suffer from friction loss and heat flows which are not single step processes. Any claim that a heat engine can beat the Carnot efficiency is called a *perpetual motion machine of the second kind* and is not feasible.

Table 3.4. Comparison of various common heat engine cycles (Spakovsky *et al* 2002).

Engine (process)	Steps	Typical efficiency (%)
Atkinson Cycle	1. Adiabatic compression 2. Isochoric heating 3. Isobaric heating 4. Adiabatic expansion 5. Isochoric cooling 6. Isobaric cooling	38%
Steam (Rankin cycle)	1. Adiabatic compression (pump) 2. Isobaric heating (boiler) 3. Adiabatic expansion (turbine) 4. Isobaric condensation (condenser)	<10% for simple engines; 25%–40% for superheated, triple expansion
Internal combustion (four stroke Otto cycle)	1. Isobaric fuel intake stroke 2. Adiabatic compression of fuel 3. Isochoric fuel ignition 4. Adiabatic expansion of burning fuel 5. Isochoric cooling 6. Isobaric exhaust stroke	25% with an upper limit of around 35% for engines made of metal
Diesel (Diesel cycle)	1. Adiabatic compression 2. Isobaric fuel ignition 3. Adiabatic expansion of burning fuel 4. Isochoric cooling	40% with an upper limit of around 55%
Sterling cycle	1. Isothermal expansion 2. Isochoric cooling 3. Isothermal compression 4. Isochoric heating	15%–30%
Gas turbine (Brayton cycle)	1. Adiabatic compression 2. Isobaric expansion 3. Adiabatic expansion 4. Isobaric cooling	48% for simple engine; up to 78% for complex designs

Six commonly used heat engine cycles are shown in table 3.4 and a brief discussion of each process follows below. The steps indicated for each cycle are for the ideal case; real engines only approximate the steps shown. The efficiencies of each of these processes, even in the idealized case, are less than the Carnot cycle efficiency. Engines using these processes have been under development for a significant amount of time. There is some room for improvement but it is unlikely that the efficiencies of real engines using these processes will improve significantly.

There are ways to use these engines more efficiently for transportation, however, a topic discussed in chapter 8.

The Atkinson cycle is slightly more efficient than the Otto cycle (below) and has been used recently in several hybrid cars, for example the Toyota Prius and the Chevrolet Volt. This cycle also describes the process of some types of rotary engines. The increased efficiency comes at the cost of lower power output but in hybrid cars an electric motor is used to increase power at peak demand (see chapter 8).

Steam engines use expanding steam to drive either a piston or turbine blades which do mechanical work. A mixture of air and water vapor is first compressed with no change in entropy (isentropic or adiabatic). The mixture is then heated at constant pressure (isobaric heating) in a boiler. The heated steam is next allowed to expand adiabatically as it pushes a piston. The final stage is to condense the vapor to a liquid at constant pressure. This process is referred to as the Rankin cycle, a chief advantage of which is that the steam can be formed from any source of thermal energy. Coal, oil, solar thermal energy, natural gas and nuclear power can all be used to generate steam for use in a steam engine.

Using 100 °C (373 K) as the hot temperature and 0 °C (273 K) as the exhaust temperature, the theoretical maximum (Carnot) efficiency of a steam engine would be $\eta = (1 - \frac{T_C}{T_H}) \times 100\% = 27\%$. Because the exhausted steam is actually at a much higher temperature than the surrounding atmospheric temperature, a simple one-stage steam engine has efficiencies of less than 10%. An improvement is to use superheated steam at higher temperatures under high pressure as the input source. It is also possible to use the exhausted steam in a second or even third piston, each operating at a slightly lower temperature (and therefore lower efficiency). A modern triple expansion steam engine using superheated steam can achieve efficiencies between 25% and 40%. It is also possible to combine the Rankine steam cycle together with a gas turbine running the Brayton cycle (described below). In a combination similar to examples of cogeneration, exhaust heat from the gas turbine is used to make steam for a steam engine giving a combined efficiency of around 40%. A steam power plant using cogeneration of heat for other purposes can capture up to 90% of the thermal energy provided by the fuel used in the plant. None of the individual steps in the process have broken the second law of thermodynamics but the efficiency of the combined process based on a definition of efficiency as the ratio of useful work output to energy input is greater than each component step.

In a modern internal combustion engine (ICE) burning gasoline is ignited by a spark from a sparkplug to form a heat source in a cycle known as the Otto cycle. The burning happens rapidly, theoretically occurring at a constant volume (*isochoric*) before the piston starts to move. The process is referred to as a four stroke cycle because, in addition to the compression and power strokes, the piston travels back and forth two extra times, first to pull the gasoline–air mixture into the cylinder and again after the power stroke to push the exhaust fumes out of the cylinder. Using the temperature of burning gasoline, 257 °C, as the hot temperature and 273 K as the exhaust temperature, the theoretical maximum (Carnot) efficiency of an ICE would be $\eta = (1 - \frac{T_C}{T_H}) \times 100\% = 48\%$. The metal of the engine restricts heat flow so that gasoline engines would quickly reach very high temperatures and melt without cooling mechanisms such as radiators. The cooling process needed to keep the engine from melting in effect introduces a two-stage heat engine which can be shown (Esposito *et al* 2010) to have a theoretical efficiency at maximum power output of

$$\eta = \left(1 - \sqrt{\frac{T_C}{T_H}}\right). \tag{3.12}$$

For our example temperatures above this bring the efficiency down to around 28% at maximum power[9]. Although various ceramic materials have been investigated that have higher melting temperatures than the metals used in typical engine construction, a ceramic engine has yet to be made commercially available.

The maximum efficiency of the ICE can also be written in terms of the compression ratio, which is the ratio of the volumes during the two isobaric steps of the process, V_1 and V_2, where V_2 is the larger volume:

$$\eta = \left(1 - \frac{Q_C}{Q_H}\right) = \left(1 - \left(\frac{V_1}{V_2}\right)^{\gamma-1}\right). \tag{3.13}$$

Here γ is the specific heat ratio, $\gamma = C_P/C_V$ where C_P is the specific heat of the gasoline–air mixture at constant pressure and C_V is the specific heat at constant volume. This equation shows that higher compression ratios give higher efficiencies. The physical strength of the piston and cylinder and the fact that gasoline will self-ignite at high pressure then become limiting factors for efficiency for an internal combustion gasoline engine.

The Diesel cycle is actually less efficient than the Otto cycle if operating at the same pressures and temperatures. However, because the diesel fuel–air mixture can be compressed to a higher pressure without prematurely igniting, the diesel engine can operate at higher compression ratios. In fact, most diesel engine designs do not require sparkplugs to ignite the air–fuel mixture, but rather the increased pressure causes the fuel mixture to spontaneously ignite. The higher compression means that the cycle encompasses more area on a P–V diagram and thus does more work per cycle than the gasoline engine. Although diesel fuel has less energy per kilogram than gasoline, diesel fuel is denser with the result that the energy per volume is about 15% higher for diesel fuel as compared to gasoline. This coupled with a higher engine efficiency means a car powered by a diesel engine typically gets as much as 40% better fuel economy than a gasoline engine. The lack of a need for an electronic ignition system and slower cycle rates of a diesel engine also make it generally more reliable and give it a longer lifetime than a gasoline engine. The diesel engine, however, also produces more carbon particulates in the exhaust which makes it a dirtier process than burning gasoline.

---

[9] This is not a maximum efficiency limit as the Carnot cycle is. Engines not running at maximum power output can have higher efficiencies.

The Sterling cycle describes a piston type engine where the combustion/heating source is external to the piston. In the two piston design a working fluid, usually air, is heated by an external source. The air expands and pushes a piston doing work. The hot piston pushes the working fluid through a heat exchanger after which it expands in a cold piston. The cold piston then compresses the fluid which ejects heat into a cold reservoir. The compressed fluid passes back through a heat exchanger to be warmed and start the cycle over. Single piston versions are also possible. The Sterling cycle is not as efficient as some other cycles but has the advantage that since the heating process is external to the piston it can be provided by a wide range of sources, for example solar thermal energy.

The Brayton or Joule cycle describes the thermodynamic process occurring in gas turbines (used in modern electricity generating plants) and jet engines. A continuous supply of burning compressed gas expands against a set of turbine blades causing them to rotate. The process is continuous and the temperature of the burning gas mixture can be much higher (more than 1000 °C) than for an ICE, which results in higher efficiencies. For a simple once-through cycle, efficiencies in real gas turbines may be as high as 48% (which is, however, lower than the Carnot efficiency for the same temperature differences). As is the case for modern steam engines, exhaust gas from a gas turbine can be recaptured for use, either for heating the intake fuel–air mixture or used with additional fuel in a second turbine (afterburner) or used as a cogeneration heat source. Multi-stage Brayton cycles can achieve efficiencies which approach the maximum Carnot efficiency. The melting temperature of the turbine blades limits the upper temperatures possible in a gas turbine or jet engine, which is a practical efficiency limit.

In 2015 the US used 38.0 EJ of primary energy sources (coal, natural gas, nuclear, hydro and wind) to make 12.6 EJ of electricity (see figure 4.5)[10]. The efficiency for these conversions, then, is about 33% which means 67% of the primary energy was converted into heat, most of which was unused. The 67% primary energy turned into heat is a direct result of the second law of thermodynamics.

3.5 Refrigerators and heat pumps

As is shown in the schematic in figure 3.3, refrigerators and heat pumps are heat engines running in reverse. A fluid with a low boiling point is allowed to expand in a process that absorbs heat from the cold reservoir. The fluid is then condensed and compressed in a process that expels heat to the hot reservoir. In the case of a home refrigerator the cold reservoir is the inside of the unit and the hot reservoir is the set of cooling coils on the back of the unit which expel heat to the surrounding air. Air

---

[10] The US Energy Information Administration (EIA) defines primary energy for combustion sources (coal, oil, natural gas, biofuels) as the caloric value of the fuel (heat given off by burning). For non-combustion sources (nuclear, hydro, solar, geothermal, wave, tidal) primary energy often refers to the secondary energy the system produces (usually net electricity produced) making a comparison between energy sources problematic. This difficulty is aggravated by the practice of some sources to list the heat energy produced by a nuclear power plant as primary energy rather than the net electricity produced.

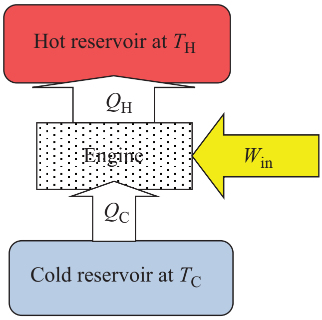


Figure 3.3. Schematic of a refrigerator. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

conditioner units operate similarly moving heat from inside the house and exhausting it to the environment from the outside unit.

In a refrigeration unit a compressed gas is allowed to expand in an arrangement of tubes which will absorb heat as predicted by the ideal gas law given in equation (3.10). The fluid is then compressed (which will also heat it) and allowed to cool in a second arrangement of tubes. The heat absorbing tubes are located in the region to be cooled (for example air inside a refrigerator) and the cooling tubes are outside the region (the air at the back of the refrigerator). This process removes heat from one place and expels it to another.

This heat transfer process can be used to heat or cool buildings in which case they are called heat pumps. The temperature of the Earth at depths deeper than about 10 m stays relatively constant in most locations and this can be used as either a cold or hot resevoir. Buried pipes are used to transfer heat from the ground during the winter to warm a building and as a heat sink for heat removed during the summer to cool the building.

Because the process involves heat transfer between two reservoirs, the efficiencies of refrigerators and heat pumps are limited by the second law in the same way as heat engines are. However in this case the efficiency, defined in equation (2.3) as $\eta = \frac{\text{energy benefit}}{\text{energy cost}}$, has a different significance. The energy benefit is heat removed from a cold reservoir, Q_C, and the energy cost is work done by a motor, W_{in}. This efficiency can obviously be larger than 1 and, to avoid confusion with heat engines where the efficiency is never larger than 1, the thermodynamic efficiency is renamed the *coefficient of performance* (COP) given as

$$\text{COP} = \frac{\text{energy benefit}}{\text{energy cost}} = \frac{Q_C}{W_{in}}. \tag{3.14}$$

A heat pump or refrigerator with a COP equal to 3 will deliver three units of energy (for example remove three units from inside a refrigerator or add three units of heat to a house) for every one unit of energy used. Since electric and gas heat turn the input work directly into heat, the COP for these sources can be no larger than one.

A heat pump used for heating with a COP less than 1 is not desirable since this would mean you would obtain more heat by just burning the fuel for heat instead of using it to run the pump.

If we assume a perfect Carnot cycle the input work will be $W_{\text{in}} = Q_{\text{H}} - Q_{\text{C}} = T_{\text{H}} - T_{\text{C}}$, in which case the maximum theoretical COP can be written as COP $= \frac{Q_{\text{C}}}{Q_{\text{H}} - Q_{\text{C}}} = \frac{T_{\text{C}}}{T_{\text{H}} - T_{\text{C}}}$. For heat engines the efficiency is larger for bigger temperature differences between the hot and cold reservoirs. For refrigerators and heat pumps the COP is larger when the temperature differences are *smaller*, which means heat pumps will be more efficient in milder climates and these efficiencies change with seasonal temperature changes. For this reason commercially sold heat pump systems are rated by the *seasonal energy efficiency ratio* (SEER) given by $\text{SEER} = \frac{\text{output cooling energy (in BTU)}}{\text{input electrical energy in the same season (in kWh)}}$. In the US the Department of Energy mandates that the minimum SEER value for heat pump systems sold in the US be 13.

Common gases used as working fluids in refrigerators and heat pumps have included CFCs (which were banned by the Montreal Protocol in 1987), hydrochlorofluorocarbons (HCFCs) and bromofluorocarbons (halons). There are hundreds of compounds which fall in these three general categories. Because the HCFCs commonly used in refrigeration are now known to block IR light from escaping the Earth, a new class of refrigerants known as hydrofluoroolefins are being phased in for air conditioning in cars, as mandated by the EPA in the US and the European Union in Europe. These compounds are slightly more flammable than earlier refrigerants but have global warming potentials hundreds of times less.

3.6 Fuel cells and batteries

All of the energy exchange processes discussed so far in this chapter are cyclic. By definition, in a cyclic process the internal energy returns to the same value at the end of the process so $\Delta U = 0$. In batteries, fuel cells and other chemical reactions, however, this is not the case. At the end of the process there have been chemical changes which leave the internal energy changed. These processes may also occur at constant volume and pressure as compared to heat engines, which do work by changing pressure to push a turbine blade or changing volume in a piston. Nevertheless there will necessarily be heat exchanges in order to compensate for changes in entropy; batteries and fuel cells get hot when they are used. Although the second law still applies to these devices the amount of expelled (or 'waste') energy is generally much less than for heat engines, so they often have higher efficiencies.

Because fuel cells and batteries involve changes in internal energy it is useful to define a quantity know as *enthalpy*, H, as

$$H \equiv U + PV. \tag{3.15}$$

One interpretation of enthalpy is the energy needed to create a system out of nothing plus the work needed to make room for the new system or, equivalently, the energy available if the system is destroyed plus the work done by the atmosphere in filling the space where the system was (Schroeder 2000). In general we are more interested in changes in enthalpy which involve changes in internal energy (chemical reactions for example) and work done on or by the system. The change in enthalpy can be measured experimentally for a given chemical reaction and standard reference tables exist for enthalpy changes in most important chemical reactions (Haynes 2016).

It is sometimes useful to subtract heat flow into or out of the system from enthalpy and define the energy needed to create a system out of nothing, *not* including energy provided by the environment from heat flow. This is called the *Gibbs free energy* and is defined as

$$G \equiv U - TS + PV = H - TS. \tag{3.16}$$

Once again we are really interested in changes in energy so for processes occurring at constant pressure and temperature we may write

$$\Delta G \equiv \Delta U - T\Delta S + P\Delta V = \Delta H - T\Delta S \tag{3.17}$$

or using the first law of thermodynamics

$$\Delta G \equiv Q + W - T\Delta S + P\Delta V. \tag{3.18}$$

Work can either be positive if work is done on the system or negative if work is done by the system. The Gibbs free energy is thus the energy available from changes in internal energy for useful work after heat flow—entropy changes and work done by expansion or compression have been accounted for. Values of ΔG can also be measured experimentally and tables exist for a wide range of chemical reactions (Haynes 2016).

Figure 3.4 shows a schematic of the chemical reactions in a fuel cell using hydrogen as a fuel. As is the case for motors or engines, energy from an input fuel is converted into work. The work done in this case is making electrical current in the form of electrons move through an external circuit. The hydrogen fuel cell is the simplest example of a fuel cell, but similar reactions can be made to occur with many other chemicals. In the hydrogen fuel cell the proton exchange membrane permits the movement of hydrogen ions through the cell without allowing the premature recombination of hydrogen and oxygen. A *flow battery* is a physical and chemical arrangement that makes the reaction reversible. In both fuel cells and flow batteries incoming material at the anode undergoes a chemical reaction which gives off electrons, while a reaction absorbing electrons occurs at the cathode. Examples of flow batteries where current from an external source can be applied to reverse the chemical reaction include zinc bromide and vanadium redox batteries. Further comments about flow batteries can be found in chapter 7.

In a hydrogen fuel cell the chemical reaction $H_2 + \frac{1}{2}O_2 \rightarrow H_2O$ occurs with the addition of an electron after it has passed through the external circuit. The enthalpy change for this reaction found in standard tables is $\Delta H = 286$ kJ, which is the energy that would come from burning a mole of hydrogen at standard temperature and pressure. This is the energy input to the system. To find the useful work possible,

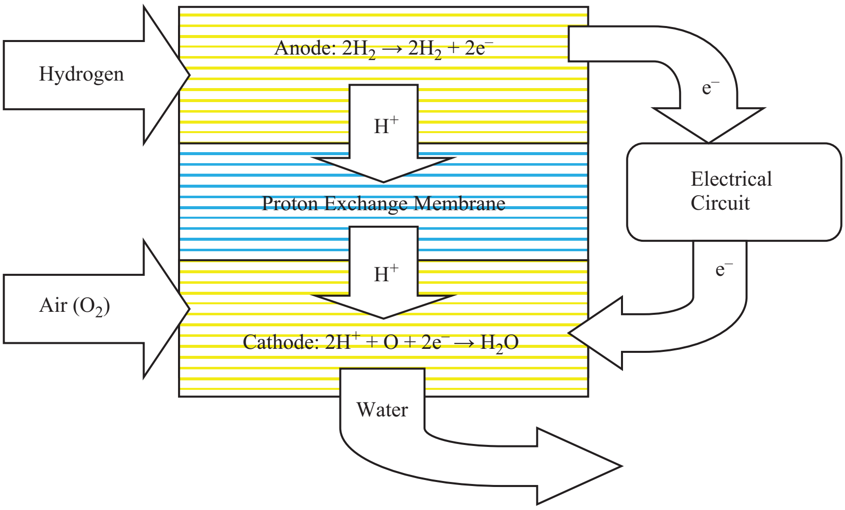


Figure 3.4. Chemical reactions of a hydrogen fuel cell. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

the heat lost due to entropy changes must be subtracted. From standard tables the entropy changes for each component of the reaction are $\Delta S_{H_2O} = -70$ J K^{-1}, $\Delta S_{H_2} = 131$ J K^{-1} and $\Delta S_{O_2} = 205$ J K^{-1}, so the total entropy change is 131 J K^{-1} + $\frac{1}{2}$205 J K^{-1} − 70 J K^{-1} = 163 J K^{-1}. Using these numbers and equation (3.18) we can calculate the Gibbs free energy, the maximum energy available to do work, which yields $\Delta G = \Delta H - T\Delta S = 286$ kJ − 300 K × 163 J K^{-1} = 237 kJ. This is the energy available from the combination of a mole of hydrogen with a mole of oxygen. The heat expelled during the process is $T\Delta S = 300$ K × 163 J K^{-1} = 49 kJ. The thermal efficiency can thus be defined as the energy available if we were to burn the hydrogen minus the equivalent heat expelled due to entropy loss divided by the energy input. For a fuel cell with these reactions we have (286 kJ − 49 kJ)/286 kJ = 83%. Hydrogen can also be burned as a fuel in a heat engine, but as we saw previously, the theoretical maximum efficiency of a heat engine is limited by the temperature at which it can operate and is considerably less than 83%. This process coupled with a high efficiency electric motor can result in much higher efficiency than a combustion engine.

Dividing the Gibbs free energy by the number of atoms involved gives the energy per reaction as 237 kJ/2 × 6.02 × 10^{23} = 1.97 × 10^{-19} J = 1.23 eV. This is also the energy per electron produced by the reaction, so the basic cell of a fuel cell using this reaction will have a voltage of 1.23 V. Higher voltages can be obtained by adding cells together in series as is done for ordinary batteries.

Many other chemical reactions and operating temperatures are possible for fuel cells. Cells have been produced using methane, ethanol and potassium hydroxide as fuels, for example. Operating temperatures range from 80 °C for hydrogen cells to 650 °C for molten carbonate cells. Efficiency is only one consideration for fuel cell applications. The heat expelled during the operation of a fuel cell and the waste products are two other considerations. Fuel cells using hydrogen have only water as a by product but other fuels generally result in carbon dioxide being given off. The application being considered will also constrain the choice of fuel cell. For example light weight applications such as computers and mobile devices should not operate at high temperatures while transportation applications demand high power and energy density (see chapter 8).

The entropy changes occurring in a battery can be calculated in a similar way as the fuel cell using standard tables for known chemical reactions. In fact a fuel cell can be thought of as a battery where the chemicals are continually resupplied, in contrast to a battery where the chemicals are limited by the initial quantity in the battery. Batteries and flow batteries will be treated in more detail in chapter 7 (on energy storage), but here we will mention that the required entropy change when a battery is being discharged turns up as expelled heat. One engineering problem involved in making a battery powered car is ensuring the heat from the battery does not build up to a dangerous level as the car accelerates. The process of charging a battery requires absorption of entropy or heat from the environment which would cool a charging battery but in most cases batteries will heat on charging and discharging due to internal resistance to current flow (equation (2.18)).

3.7 Entropy changes for electrical processes

A current carrying wire in a magnetic field will experience a force acting on it called the Lorentz force

$$\mathrm{d}\vec{F} = I\vec{dl} \times \vec{B}, \tag{3.19}$$

where $\mathrm{d}\vec{F}$ is the force in Newtons on a small length of wire, $\vec{dl}$, in meters carrying current, I, in amperes. The magnetic field, $\vec{B}$, is in tesla and the cross product of the two vectors tells us that the force acts perpendicular to both the magnetic field and the wire. Because of this force, a loop of wire will experience forces so that the loop will turn, as shown in figure 3.5.

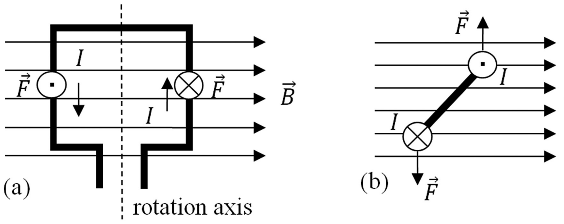


Figure 3.5. A loop of wire carrying current I in a magnetic field, $\vec{B}$. In (a) the current goes up the wire on the right producing a force into the page and down the wire on the left producing a force out of the page. The top is parallel to the field and so no force is produced on the loop. (b) shows a top view, the current comes out of the page on the right producing a force upward and into the page on the left producing a force downward. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

If the directions of current and magnetic field remained fixed the loop would rotate to a position where the forces were opposite and the loop would oscillate back and forth until friction brought it to a standstill. By a clever use of sliding connectors, called *brushes* and *commutators*, the direction of current can be reversed every half turn so the force changes direction and keeps the loop rotating. An alternative is to reverse the direction of the magnetic field each half turn which can be done using alternating current and electromagnets. Overlaying many loops increases the force on the configuration so *electric motors* typically have large coils of rotating wires. Three coils oriented at 60° to each other are often used so that there is no spot at which the force is zero as the loop turns.

Faraday's law (equation (2.19)) says a voltage appears in a circuit or wire as the result of the change in magnetic flux. As an application of Faraday's law, if a wire is forced to move through a magnetic field so that it sweeps out an area perpendicular to the field, an electromotive force (emf) ε, in volts will appear across the wire:

$$\varepsilon = Blv, \tag{3.20}$$

where l is the length of wire moving at velocity v perpendicularly through a magnetic field B. The arrangement in figure 3.5 can therefore be used in reverse to get electric current out using input mechanical energy. If the loop, with no current initially flowing, is forced to turn in the magnetic field, an emf (in volts) is generated. If the loop is connected to an external circuit, a current will flow. This is the basis behind an *electric generator* and, in fact, in many early applications of electrical circuits, generators were simply motors being run in reverse (the three coils in a generator at 60° to each other produce *three phase electric power*, the common type of electricity supply).

The first law of thermodynamics still applies for the case of a generator. It is harder to turn the loop if current is produced and the electrical energy output is no larger than the mechanical input energy. The gas mileage of an automobile goes down slightly when electrical devices such as lights and radio are used because the gasoline motor has to work harder to create the current flow needed to recharge the battery. Any scheme attempting to use an electric motor to turn an electric generator which then runs the motor plus other devices will not work; the output energy in the current is proportional to the applied mechanical work. Even without friction losses the scheme would at best produce no more energy than was originally input.

Unlike heat engines, the limits to efficiency of motors and generators due to the second law of thermodynamics are exceedingly small. An ideal motor with no friction or other loss can have a theoretical efficiency of better than 99% and real electric motors have been built with efficiencies close to this limit. The second law limit to efficiency for electrical processes has to do with the power needed to overcome random thermal energy of the electrons already in the wire so that current will flow. It can be shown that this power is on the order of

$$P = \frac{V^2}{R} = 4k_{\mathrm{B}}Tb, \tag{3.21}$$

where k_{B} is Boltzmann's constant and T is temperature (Nyquist 1928). Here b is the bandwidth, the range or variation of frequencies in hertz of the alternating current being used in the circuit. Because Boltzmann's constant is very small, the power needed to overcome thermal resistance is practically negligible with the result that an ideal electric motor can have near perfect second law efficiencies.

For real electric motors there are mechanical friction losses as well as resistance losses proportional to $P = I^2R$, as discussed in chapter 2. These losses are *not* limitations due to the second law of thermodynamics. Decreasing the resistance by making the wires thicker (or potentially superconducting wires) will increase the efficiency but at a higher economic cost and greater weight[11]. For stationary motors where weight is not a factor, efficiencies can be very high. The fluctuating magnetic fields involved in the motor are also not perfectly confined to the motor, with the result that electromagnetic waves radiate into the surroundings and this also constitutes lost energy. Well-designed low horsepower electric motors (<1000 W) typically have efficiencies around 80% while larger motors (>95 kW) have efficiencies as high as 95%. These efficiencies are significantly higher than the efficiencies of fuel cells and any heat engine we have discussed. The theoretical, second law limitation to efficiency is extremely small for electrical processes. As we will see in chapter 8, this has important implications for transportation choices in the future.

Projects

1. Write a report about and find online simulations of P–V and S–T diagrams of the various heat engines shown in table 3.4. Include a thorough description of each and include examples of where they are used. What are the pros and cons of each type of engine? What are their efficiencies in practical applications?
2. Find the amount of pollution and greenhouse gas per output work given off by existing versions of each of the engines shown in table 3.4. Which engines are cleaner? Which are closer to their theoretical maximum efficiency?
3. Diesel cars are more popular in Europe than in the US. Report on the pros and cons of each type of motor (diesel and petrol), their efficiency, their costs (including fuel) and the pollution each generates, and generate an argument for why there is this difference in popularity. Include a discussion of fuel taxes.
4. Several cycles of existing motors are not shown in table 3.4, and there are many versions of the ones shown. Report on as many of these other cycles as you can find, describing each and their practical applications. Why was the Atkinson cycle chosen for the Prius?
5. Explain how the following processes do not break the first or second laws of thermodynamics: a bird egg becomes a chick, water freezes into ice cubes in your freezer, crystals form from a salt solution that evaporates, the drinking bird toy bobs up and down for hours (www.youtube.com/watch?v=gZ-zeRCUicw).

---

[11] Electric motors using superconducting wires are still in the development stage.

6. Find details about the most efficient heat pumps available today and report on what you find. Include descriptions, efficiencies, applications, etc.
7. Write a summary of how the following refrigeration processes work: gas-powered absorption refrigerator, thermo-acoustic cooling, thermoelectric cooling, Einstein–Szilárd refrigerator (there are several designs, some of which are being used in various tropical and desert climates with solar energy as the energy source) and tri-generation (similar to co-generation).
8. Investigate the chemical reactions occurring in other kinds of batteries (for example lithium ion, metal hydride, lead storage) and fuel cells which are now under investigation for possible commercial applications. Calculate their theoretical efficiencies and cell voltages following the example in the text.
9. Write a report on the current status of the development of flow batteries. Include calculations of the chemicals being used and their efficiencies.
10. It is often very difficult to locate the flaw in an inventor's claim that a heat engine can beat the Carnot efficiency maximum. Find examples on the Internet of claims that either the first or second law has been broken. Analyze a few of these schemes and comment on whether the processes shown might be possible or not.

References

EERE (Office of Energy Efficiency and Renewable Energy) 2016 Types of fuel cells http://energy.gov/eere/fuelcells/types-fuel-cells (Accessed: 17 January 2017)

Esposito M, Kawai R, Lindenberg K and Van den Broeck C 2010 Efficiency at maximum power of low-dissipation Carnot engines *Phys. Rev. Lett.* **105** 150603

Forinash 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Glazer M and Wark J 2001 *Statistical Mechanics, A Survival Guide* (Oxford: Oxford University Press)

Haynes WM (ed) 2016 *CRC Handbook of Chemistry and Physics* 97th edn (Boca Raton, FL: CRC Press)

Nyquist H 1928 Thermal agitation of electric charge in conductors *Phys. Rev.* **32** 110

Schroeder DV 2000 *Thermal Physics* (New York: Addison-Wesley Longman)

Spakovsky ZS, Griezer EM and Waitz IA 2002 Thermodynamics and propulsion notes http://web.mit.edu/16.unified/www/SPRING/propulsion/notes/notes.html (Accessed: 17 January 2017)

Chapter 4

Non-renewable energy

Changes in global primary energy use over the past 35 years are shown in figure 4.1. Global energy consumption was about 146 EJ (EJ = exajoule = 10^{18} J = 23.9×10^6 metric tonne oil equivalent) in 1960 but has increased to 552 EJ in 2012, reflecting both the doubling of the population and the increase in standard of living mentioned in chapter 1. Petroleum consumption has steadily increased in this time frame, coal consumption increased a little faster after 2000 than before due to demands from a growing economy in China and India. Nuclear energy has seen a slight drop due to the closing of ageing reactors and political reactions to the Fukushima disaster, and biofuels are increasing in use, as can be seen in figure 4.1. Global *energy intensity*, a ratio of the total energy used divided by the gross world product, declined in this same time period indicating a more efficient use of energy to produce wealth (Ma 2014). Although we are using energy more efficiently to produce growth and wealth, the amount we are using is increasing. This chapter looks at historical changes in energy use, sources of fossil fuels and projections of future supplies.

4.1 Fossil fuels

In 2014 some 31% of the global energy supply was from petroleum, 28% from coal, 21% from natural gas, 14% from renewables and 5% from nuclear energy (IEA 2016). In other words 80% of the world's energy comes from fossil fuels. The energy content of a liter of gasoline is about 34.6 MJ when burned; using 552 EJ per year is equivalent to burning 5.05×10^5 liters of gasoline *per second*. As seen in figure 4.1, renewable energy (still mostly hydroelectric) and biofuel use have increased in recent years, but it will be a very long time until fossil fuels are replaced by other sources.

Energy use is not distributed evenly around the globe. Figure 4.2 shows the energy use in different regions of the world along with the percentage of the population (in parentheses). North America consumes 22% of the world's energy but has only 5% of the global population. In 2000 Asia used about the same amount of

doi:10.1088/978-1-6817-4493-3ch4

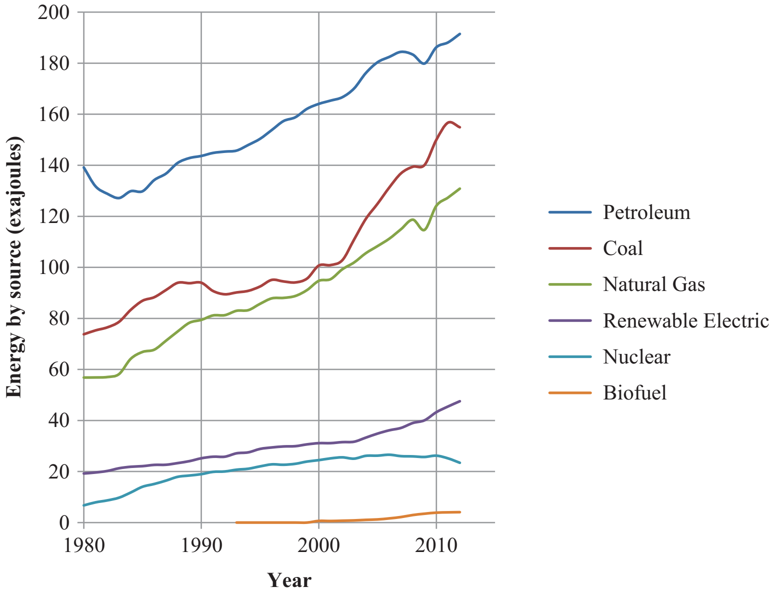


Figure 4.1. Global energy consumption by source (EIA 2016a, BP 2016). Renewables: 35% hydroelectric, 22% wood, 21% biofuels, 13% wind, 5% burning waste, 2% geothermal and 2% PV.

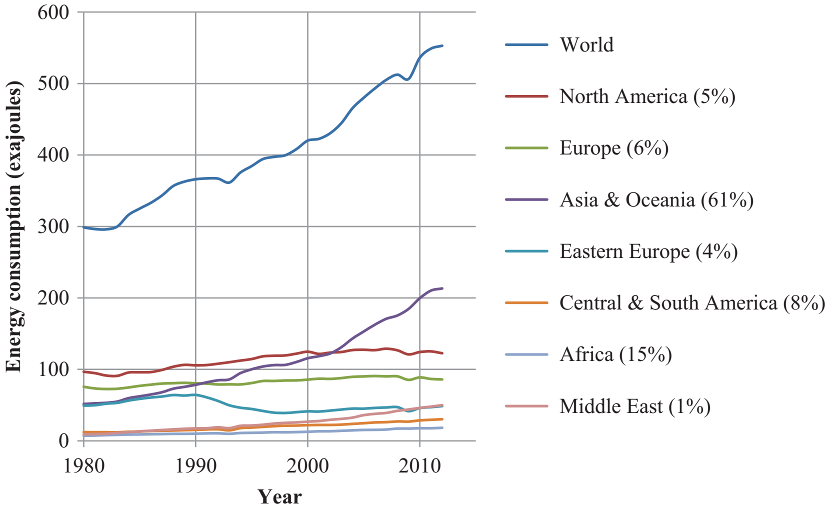


Figure 4.2. Global energy use and percent of the global population by region. The downturn for Eastern Europe after 1990 reflects the breakup of the Soviet Union. The steady increase in Asia reflects the recent strong economies of China and India (EIA 2012).

energy as North America but had 60% of the population. The economies of India and China grew very strongly between 2000 and now, resulting in a much higher standard of living and a parallel jump in energy consumption. If the rest of the world reaches a per capita energy use equivalent to North America or Europe the world will clearly need a much larger supply of energy and in fact the US EIA projects there will be a global increase in energy demand of 56% by 2040.

It is not clear if the developing world will follow in the footsteps of the developed world in terms of exploiting various energy sources, but it is interesting to look at the historical energy use in the US as an indication of what might happen. Prior to 1850 nearly all the energy used in the US (and the rest of the world) came from wood and other biofuels such as corn stover (figure 4.3). That trend decreased in the 1950s but is seeing an uptick recently because of government subsidies for bio-fuels in the US. Coal was the next energy source to undergo significant development followed by petroleum starting in 1900. Growth in petroleum and natural gas use largely offset increases in coal consumption until the late 1970s when petroleum consumption started to decrease. The decline in petroleum use was compensated by an increase in coal and nuclear sources after 1970. Hydroelectric leveled off after 1975 because most of the easier to tap resources were dammed and large government funded infrastructure projects were dropped from the national agenda in the US and many other countries. Nuclear energy has also leveled off in the past few decades because of historically strong public opposition to new nuclear power plants, although nuclear still supplies 19% of the electricity in the US. Geothermal, wind and PV (photo-voltaic; see the next chapter) sources make up an insignificant part of the US energy supply, although wind has increased sharply since 2002 in some locations, currently contributing a little more than 5% of the electricity generated.

An indication of the multifaceted relationship between energy consumption, economics and global politics can be gleaned from connecting some historical facts with the US energy consumption shown figure 4.4. Starting around 1970 there was a growing gap between production in the US and consumption, necessitating an increase in energy imports. This shortfall occurred globally (although the US was by far the world’s largest energy consumer) and was met largely in the form of petroleum exported by the Middle East. This increase in the importance of global petroleum sources marks the beginning of a strong political interest of the US in Middle Eastern politics. The dip in energy use starting in the early 1970s was

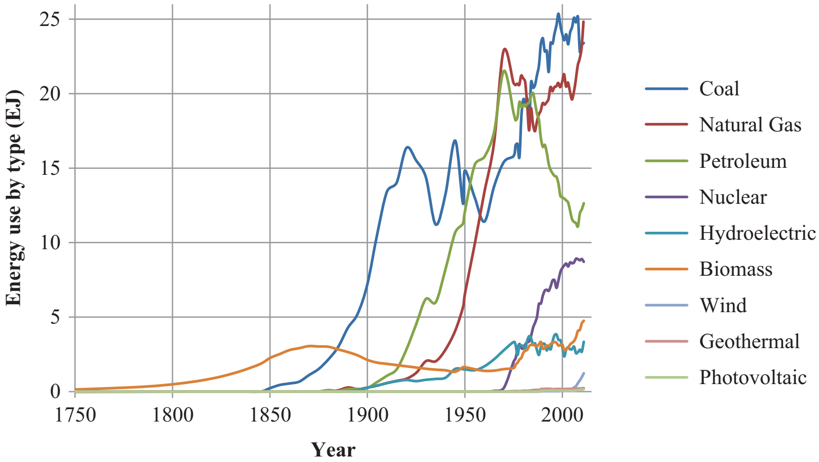


Figure 4.3. Historical energy use in the US by source (BP 2016).

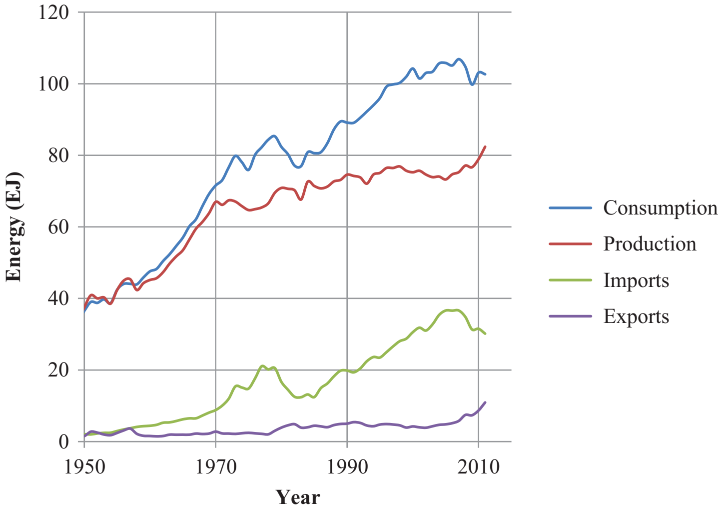


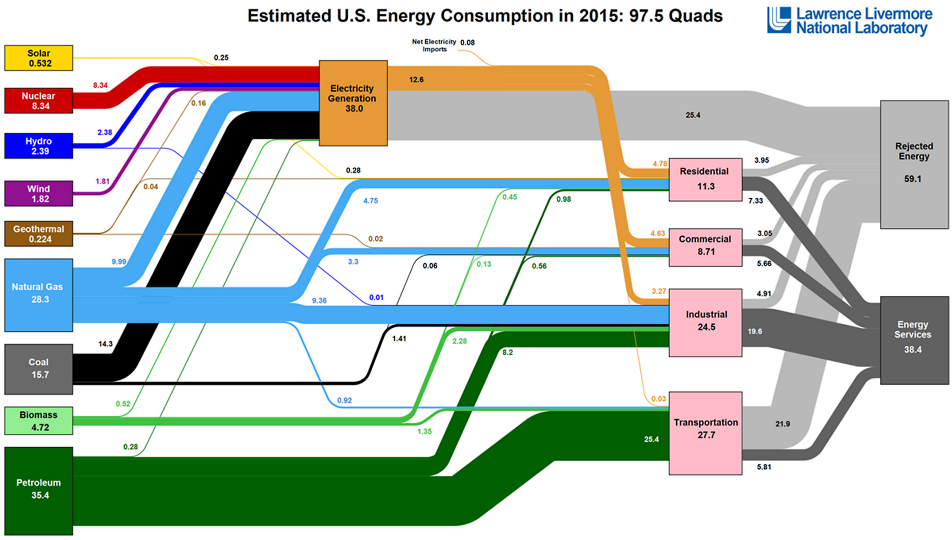
Figure 4.4. Total energy use in the US historically (EIA 2012).

partially due to an OPEC (Organization of the Petroleum Exporting Countries) oil embargo in reaction to American involvement in the Yom Kippur War and a subsequent economic downturn. By the late 1970s conservation measures such as a lower maximum highway speed were implemented by then President Jimmy Carter. A slight downturn in energy consumption can be seen in 2008 during the global recession of that year. Imports declined and production increased slightly since 2005 due to more oil and natural gas on the market provided by hydraulic fracturing. The decrease in coal and sharp increase in natural gas production after 2006 reflects both a lowering in the price of natural gas due to fracking and the prospect of future political pressure to reduce emissions contributing to climate change[1]. It should be clear from this very brief foray into historical influences that energy consumption has a complex, codependent connection with economics and the global political environment.

One other interesting aspect of figure 4.4 concerns the exported energy shown. Why would a country export energy at the same time it was having trouble meeting its consumption needs? This situation occurs because energy, particularly petroleum, is traded on a global market. Most of the oil imported to the US actually comes from Canada (the US imports about one third as much oil from Saudi Arabia as it does from Canada) but this would not be possible if the Middle East did not supply other parts of the world with oil. The US exports petroleum and petroleum products to much of the rest of Central and South America because it is closer to the US than the Middle East. There is an economic incentive for petroleum to travel the shortest distance from source to consumer and this results in exports as well as imports occurring at the same time. The amount of imports into the US, however, far outweigh the exports.

---

[1] Natural gas produces 2.75 kg of CO_2 per kg burned while coal produces 3.67 kg CO_2 per kg of coal burned.



Source: LLNL March, 2016. Data is based on DOE/EIA MER (2015). If this information or a reproduction of it is used, credit must be given to the Lawrence Livermore National Laboratory and the Department of Energy, under whose auspices the work was performed. Distributed electricity represents only retail electricity sales and does not include self-generation. EIA reports consumption of renewable resources (i.e., hydro, wind, geothermal and solar) for electricity in BTU-equivalent values by assuming a typical fossil fuel plant heat rate. The efficiency of electricity production is calculated as the total retail electricity delivered divided by the primary energy input into electricity generation. End use efficiency is estimated as 65% for the residential sector, 65% for the commercial sector, 80% for the industrial sector, and 21% for the transportation sector. Totals may not equal sum of components due to independent Rounding. LLNL-MI-410527

Figure 4.5. US energy consumption, source and use. 1 quad is a quadrillion BTUs which is equal to 1.055 EJ.

Figure 4.5 shows the sources of primary energy in quads (1 quad = 10^{15} BTU = 1.055 EJ) for the US and how this energy is used[2]. Other developed countries probably have similar distributions of energy flow. Some electricity is provided from solar, nuclear, hydro, wind and geothermal, but most comes from burning coal and natural gas. Of the 38 quad of primary energy used to make electricity, 25.4 quad is turned into waste heat, mostly because of the second law of thermodynamics, representing an efficiency of 33%. Almost none of this electricity is used for transportation; most is used for residential and commercial purposes. Transportation uses about 27.7 quad, 28% of the total energy consumed, nearly all coming from petroleum. The net transportation efficiency is about 21%, again severely restricted by second law losses from using the ICE (internal combustion engine). Industrial process are the most efficient with only 20% lost to waste heat. As mentioned previously, industrial processes are often coupled so that heat loss from power generation is used for other purposes. The net efficiency of the US energy cycle in 2015 was 39.4% in converting primary energy into useful benefit.

Table 4.1 puts the US energy consumption by sector seen in figure 4.5 in perspective by comparing it with a few other select countries and regions. The developed world (Organization for Economic Cooperation and Development (OECD)) uses about 24%

---

[2] As mentioned in chapter 3, the US Energy Information Administration (EIA) defines primary energy for combustion sources (coal, oil, natural gas, biofuels) as the caloric value of the fuel (heat given off by burning). For non-combustion sources (nuclear, hydro, solar, geothermal, wave, tidal) primary energy often refers to the secondary energy the system produces (usually net electricity produced) making a comparison between energy sources problematic. This is aggravated by the practice of some sources to list the heat energy produced by a nuclear power plant as primary energy rather than the net electricity produced.

Table 4.1. Energy use by sector for selected regions in 2012. OECD members include the US, Canada, Mexico, most of Western Europe, Japan, South Korea, Australia and New Zealand. The numbers do not add up across the rows because electrical use is calculated separately and includes use in other categories (EIA 2016a). 1 quad is a quadrillion BTUs which is equal to 1.055 EJ.

Entity	Residential	Commercial	Industrial	Transportation	Electricity	Total quad	Per capita ($\times 10^9$ J yr^{-1})
World	53.0	29.3	222.3	104.2	210.7	549.3	77.7
OECD	26.6	20.0	73.3	57.5	93.8	238.5	191.9
Non-OECD	26.4	9.3	149.1	46.8	116.9	310.8	53.3
US	10.3	8.2	24	26.2	38.3	94.4	299.7
China	8.2	2.5	59.0	12.8	50.3	115.0	83.3
India	2.1	0.6	13.1	3.3	10.8	26.2	21.1
Africa	1.8	0.5	10.0	4.3	7.0	21.5	20.9
Brazil	0.8	0.5	7.5	2.9	5.0	15.2	76.0
Other non-OECD Americas	1.3	0.5	6.3	3.6	5.3	15.8	57.9

of its energy for transportation compared to the US which uses 28%, China using 11%, India using 13% and Brazil using 19%. The less developed part of the world uses 15% of its energy for transportation on average.

The developed world uses 31% of its energy for industrial purposes while the developing world uses 48% of its energy in industry. Part of this disparity is that developing countries have a lower standard of living so energy used for residential and transportation per capita is much lower compared to the developed world[3]. A second factor is that a large fraction of the industrial processes that were first developed in the OECD countries has been outsourced to developing countries where labor is cheap. For example, China currently makes 90% of all computers, 80% of all air conditioners, 60% of all shoes, and 70% of all mobile phones manufactured globally (Economist 2015).

The per capita energy use in the US is one of the highest in the world at nearly 300×10^9 J per year, one and a half times the average for the developed world, nearly six times the average for the less developed world, three and a half times the average for China and fourteen times the energy used per person in Africa and India. This is in part due to higher transportation energy use, but also because houses are much larger there, requiring more energy to heat and cool, and electronic device (computers, tablets, mobile phones, music devices, etc) ownership is more widespread.

4.2 Petroleum

The most studied fossil fuel is petroleum or crude oil. Petroleum was formed millions of years ago from organic material originating from marine life. When these ancient

---

[3] Approximately 1.5 billion people in the world live with no electricity and 1.3 billion use traditional fuels such as wood, straw and dung for heating and cooking (IPCC 2012).

marine organisms died they sank and were covered over with other sediment or folded into the Earth's crust by tectonic activity. As the buried material was pushed further down into the Earth's crust it was heated by the Earth's core which remains molten due to radioactive heating. Under anaerobic conditions organic material will undergo a chemical process called *pyrolysis* when heated between 50 °C and 120 °C to form oil and natural gas. The longer the heating process is maintained, the more natural gas is formed. Most petroleum deposits also contain natural gas and the ratio is determined by the length of time the material is buried and how close it is to the Earth's core. As the depth of this layer increases the temperature and pressure also increase, so that eventually the oil and natural gas fracture the surrounding rock. At this time they migrate upwards through the fractured rock until they reach an impermeable boundary to collect as conventional oil, or the oil adheres to sand or shale[4] where it is found as shale oil or bitumen. Once they have escaped the heat and pressure from deeper in the Earth's crust, pyrolysis stops and the oil and natural gas are relatively stable. Petroleum reaching the surface of the Earth is broken down by bacteria which will use the hydrocarbon compounds in the oil as an energy source.

About 43% of a barrel of oil is made into gasoline, 22% becomes diesel fuel, 9% becomes jet fuel, 15% is used by the petrochemical industry and the remaining 11% becomes heavy fuel oil, liquefied petroleum gases (LPGs) and other distillates used as lubricants. These products evaporate at different temperatures and so can be separated from crude oil by heating; a process called *fractionation*. Petrochemical industry products include plastics such as acrylic fiber, acrylonitrile butadiene styrene (ABS), polyvinyl chloride (PVC), polyethylene (PE), polyethylene terephthalate (PET), polypropylene (PP) and polystyrene (PS). Other industrial products based on chemicals extracted from petroleum include adhesives, resins, inks, specialized coatings, cosmetics, food additives, pharmaceutical drugs and basic stock chemicals such as benzene, propylene and methanol. Even if we did not use petroleum as fuel it would remain an important resource for the manufacturing industry.

It is important to have some understanding of the technical terms used in the literature when describing petroleum resources. *Oil in place* refers to an estimate of how much oil is believed to be in the ground based on known geological features, data from *seismology*, *gravimetric* and *magnetometer* measurements, as well as sample drilling. A resource is termed a *proven* reserve if there is an 80% to 90% confidence, usually based on test wells, that this amount can be extracted using current extraction techniques. Proven reserves can be further broken down into developed reserves (fields which already have wells) and undeveloped reserves. *Unproven* reserves may be divided into *possible* reserves (with 10% confidence that oil can be extracted using current techniques) and *probable* reserves (50% confidence). Until very recently none of these estimates included non-conventional oil (sometimes called tight oil) such as shale oil or tar sands because extraction technologies were not well developed. The *estimated ultimately recoverable* petroleum is the cumulative oil production to date (the oil that has already been extracted) plus the proven reserves. Similar terminology

---

[4] Shale is a layered sedimentary rock formed from fine mineral particles found in mud which has been compressed and heated.

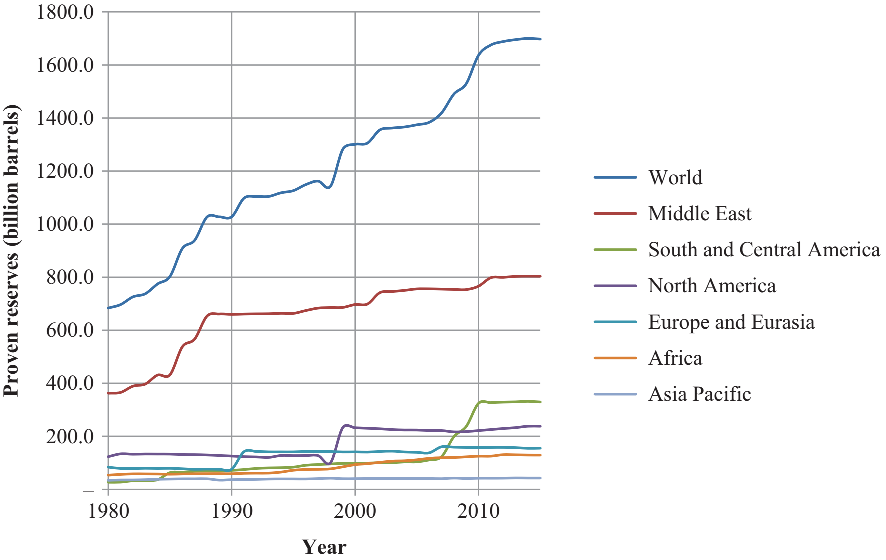


Figure 4.6. Historical changes in proven oil reserves by region (BP 2016).

applies to natural gas, coal, shale oil and tar sand deposits. A chart of historical proven oil reserves is shown in figure 4.6.

It is important to realize that proven reserve figures have changed over time both for political reasons and as extracting technology has improved. Part of the sudden increase in proven reserves in North America shown in figure 4.6 is due to the fact that Canada started counting the huge resources of the Athabascan oil sands as proved reserves. Technology made exploitation of this resource economically feasible so Canada's proven reserves increased from a few tens of billions of barrels of oil to over 1000 billion barrels (Gray *et al* 2004). In the late 1980s several countries in the Middle East re-evaluated how much petroleum was believed to be held in the oil fields there, resulting in the increase seen in figure 4.6. The estimated amount of petroleum in an oil field is usually proprietary information known only by the private company doing the extraction and the country that hires it, and there are political and economic reasons for both parties to over-estimate this number. As a result it is not always clear if all increases in proven reserves reflect real increases in the oil in place. The question 'how much oil is there in the world?' does not have a simple answer and the answer changes over time. Similar problems arise when evaluating other fossil fuel resources.

The most easily accessed petroleum, known as *conventional* oil, is found trapped in porous rock formations capped by an impermeable layer which prevents it from reaching the Earth's surface. Oil in this form is generally found under pressure, layered between natural gas, which is lighter, and very saline water, which is denser. Oil from these formations can be extracted simply by drilling through the impermeable cap and the oil will initially flow with little or no pumping. This *primary extraction* technique taps about 20% of the oil in the reservoir. Once the naturally occurring pressure drops to the point that the well is not producing efficiently, another 15% of the oil in the reservoir can be extracted with *secondary techniques* such as pumping or pressurizing the oil by injecting natural gas, water, air or other gases. *Tertiary extraction* techniques include heating the oil in the reservoir to make it less viscous either by injecting steam or burning some of the oil in the reservoir which then heats the remaining oil. Detergents can also be injected which make the oil less viscous and easier to pump. Generally 40% of the oil in the reservoir is left after these technologies, depending of the viscosity of the petroleum which varies depending on the exact chemical composition. This remaining oil may eventually be accessible by new technologies.

Petroleum may also be found trapped in layers of shale, called *shale oil* or mixed with sand or clay which is termed *tar sand* or *bitumen*. These resources cannot easily be pumped out of the ground because the oil adheres to the rock or sand due to surface tension. Until recently the only way to extract these resources was by crushing and heating the shale or tar sand either in place or after it was dug up. Shale deposits are relatively thin, a few tens of meters thick and usually found a few kilometers deep, so they were very difficult to access. More recently it has become possible to extract this resource using hydraulic fracturing (fracking). Fracking was developed in the 1950s but only began to be used on a large scale after 2000 for the extraction of oil and gas from shale oil deposits. A well is drilled down to the shale layer and then turned to drill horizontally into the layer. Water and other chemicals are pumped into the well at high pressure to fracture the shale layer, releasing the trapped oil and gas which can then be pumped out, once the water has been removed.

In theory fracking should not interfere with drinking water supplies because it occurs at depths up to several kilometers below the water table. However there is anecdotal evidence that some contamination can occur. The speculation is that old existing wells (possibly unidentified) and leaks in the casing of new wells could lead to contamination of ground water. The water pumped into a fracking well has proprietary chemicals added to it to enhance the oil extraction and returns to the surface with these chemicals as well as contaminants from the ground such as arsenic and radioactive isotopes. This waste water is often held on the surface in large ponds or injected back into a well after the oil and gas has been extracted. To date the only verifiable groundwater contamination has been from contaminated waste water used in the process being stored in surface ponds which have leaked.

A second consequence of fracking has now been verified and has to do with earthquake activity. Prior to 2009 the average number of earthquakes larger than magnitude three occurring in the central part of the US was fewer than seven per year. The number jumped to between 75 and 190 per year from 2011 to 2013 and was more than 650 in 2014 (Weingarten *et al* 2015). Waste water is often injected into a well after the resource has been extracted to get rid of the contaminated residues and the injection rate of this process correlates to the surplus number of earthquakes measured. The correlation is stronger in zones where there is a potential for earthquake activity due to geological structure. Fortunately the earthquakes tend to be small in magnitude but several have been above magnitude five on the Richter scale.

Extraction from tar sand generally starts by surface mining the clay or sand and then mixing it with hot water and air. The petroleum particles in the sand adhere to air droplets in the mixture and rise to the top where they can be separated from the remaining material. This method requires significant amounts of energy for heating the water that is used. The waste, including water left after the process generally has a volume 35% larger than the extracted material which makes it difficult to dispose of as the leftover debris cannot just be put back into the location where the tar sand was removed because it has a larger volume and is in the form of a semi-liquid sludge.

Based on known reserves and the rate of use, a crude estimation can be made of how much longer oil reserves might last (a more sophisticated calculation is given below). Global production of petroleum in 2015 was 91 670 thousand barrels a day, while consumption that year was 96 008 thousand barrels a day. The annual rate of consumption represents about 2.1% of the global proven reserve. Assuming the same rate of consumption and no changes to the proven reserves there would be about 48 years of petroleum left.

4.2.1 Darcy's law

Darcy's law is an empirically derived equation governing the flow of liquids and gasses through a porous solid, including water in an aquifer and petroleum through an oil field. The rate of flow, Q, in $\text{m}^3\,\text{s}^{-1}$, of the liquid is proportional to the pressure gradient, the cross sectional area of the flow, A, and the permeability of the solid, κ in m^2. The flow rate is inversely proportional to the dynamic viscosity, μ of the liquid in $\text{kg}(\text{s}\cdot\text{m})^{-1} = \text{Pa}\cdot\text{s}$ and the distance the fluid will travel, L. Darcy's law for flow in one dimension is

$$Q = -\frac{\kappa}{\mu}\frac{A\Delta P}{L}, \tag{4.1}$$

where the pressure gradient, ΔP, is the difference in pressure between two locations, for example the pressure in the oil field at one location and the surface pressure at the location where an extraction well is to be drilled. This equation can also be used to determine the rate at which a pollutant travels through the soil.

The measure of the ability of a solid to allow fluids to flow through it is termed *permeability*. Gravel and loose sand have permeabilities near $1 \times 10^{-8}\,\text{m}^2$ whereas solid granite has a permeability of around $1 \times 10^{-18}\,\text{m}^2$. Oil field permeabilities typically range from $1 \times 10^{-11}\,\text{m}^2$ to $1 \times 10^{-13}\,\text{m}^2$. Dynamic viscosity is a parameter indicating how easily an incompressible fluid will flow and is temperature dependent. The viscosity of water at room temperature is approximately $1 \times 10^{-3}\,\text{Pa}\cdot\text{s}$ and varies by a factor of 10 between freezing and boiling temperatures. The dynamic viscosity for crude oil varies between $5 \times 10^{-3}\,\text{Pa}\cdot\text{s}$ and $20 \times 10^{-3}\,\text{Pa}\cdot\text{s}$, depending on the exact chemical makeup of the oil and its temperature. The degree of saturation of the solid by the liquid is another factor involved in the flow rate. Porous solids which are not completely saturated exhibit a resistance to fluid flowing through them because of capillary action; because of liquid surface tension the fluid clings to the surfaces of the channels through which it is flowing. This additional resistance to flow is sometimes referred to as solid–liquid suction. The effect can be modeled as a decrease in the permeability of the unsaturated solid.

4.3 Natural gas

Natural gas is mostly methane, CH_4 but can contain up to 5% of other carbon compounds. Methane is a relatively clean energy source because, unlike petroleum or coal, it is almost pure carbon and hydrogen. Additionally the hydrogen provides energy when methane is oxidized forming H_2O without creating extra carbon dioxide. As mentioned previously, natural gas forms the same way oil does and is usually found in association with oil deposits. In the early days of the oil industry natural gas was considered a nuisance and was simply burned off because there was not a convenient way to store or transport it.

Transporting natural gas is more difficult than transporting petroleum with the result that natural gas is usually extracted and used locally, rather than being imported. It is questionable that there will be a significant increase of natural gas imported to the US from resources such as the Middle East because of the difficulty of transportation, in contrast to Europe which is close enough for pipeline distribution or delivery by liquid gas tankers. The most efficient form to transport natural gas if a pipeline is not feasible is to cool the gas to −160 °C, the temperature at which it liquefies into liquid natural gas, but there is a greater potential for explosion in the event of a tank rupture or leak than for petroleum.

As can be seen in figure 4.7, the amount of proven natural gas reserve has changed over time. As was the case for petroleum, not all of these changes are the result of newly discovered resources; technology such as hydraulic fracturing have had a great impact on proven reserves. In 2015 production of natural gas was 3538.6

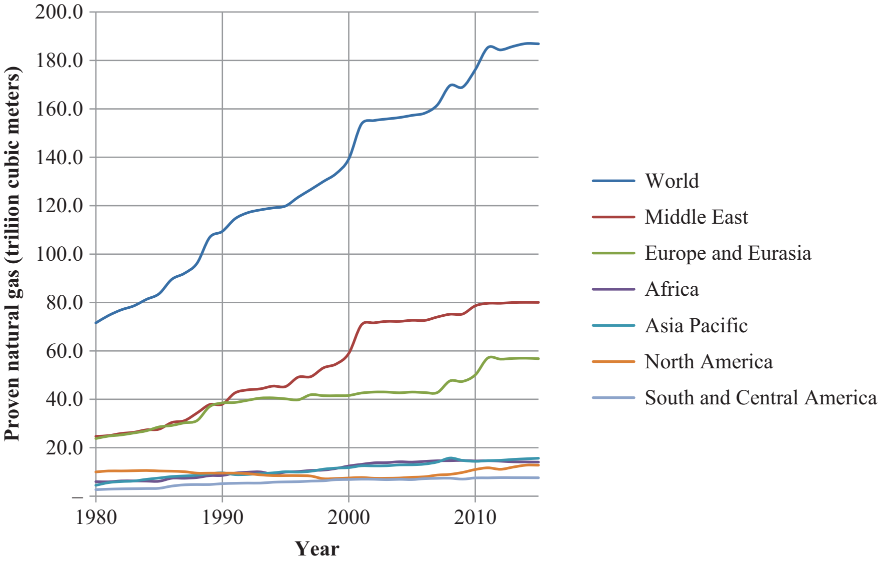


Figure 4.7. Historical changes in proven global natural gas reserves (BP 2016).

billion cubic meters and consumption 3468.6 billion cubic meters, representing about 4% of the proven reserve. If no new natural gas is found there would be about 25 years of natural gas left, assuming the rate of consumption does not change.

4.4 Coal

Coal forms from a process very similar to oil and natural gas but starting with terrestrial plant material. Accumulation of decayed material in bogs, called peat, become a carbon sink when they are covered over by other inorganic material. Millions of years ago large amounts of peat were buried as the result of tectonic activity and heated anaerobically to high temperatures forming coal. Because coal forms from solid material, fossilized remains of plants and animals can often be found in coal, which was one of the first indications that coal was formed in the distant past and not more recently. Once the coal is formed, continual crustal folding eventually brings the coal deposit closer to the surface where it can be extracted as a fuel source. The carbon and other compounds found in coal vary a lot more than in oil or natural gas and depend on when and exactly how it formed. The oldest type of coal is *anthracite*, which formed around 360 million years ago and has a carbon content of around 95%. *Bituminous* coal has a carbon content between 50% and 80% and formed some 300 million years ago. Lignite, the youngest coal, formed about 150 million years ago and has a carbon content less than 50%. Burning any of these types of coal produces carbon dioxide and other pollutants such as sulfur, but because anthracite has the highest carbon concentration it is the cleanest burning.

The US, Russia and China have by far the largest proven coal reserves, which is evident in the values in table 4.2. Currently coal consumption is decreasing in the US as old coal driven electric power plants are being retired and new natural gas plants are being built to replace them. Coal consumption in China, however, is increasing sharply as new coal fired plants are being built to supply the growing energy needs of a rising economy.

Global coal production in 2015 was 7861.1 million tonnes while consumption was 3839.9 million tonnes oil equivalent, which represents 2641.8 million tonnes of coal. Annual consumption represents 0.3% of the proven reserve so that coal reserves

Table 4.2. Global proven coal reserves by region in 2015 (BP 2016).

Million tonnes	Anthracite and bituminous	Sub-bituminous and lignite	Total
North America	112 835	132 253	245 088
South and Central America	7282	7359	14 641
Europe and Eurasia	92 557	217 981	310 538
Middle East and Africa	32 722	214	32 936
Asia Pacific	157 803	130 525	288 328
World	406 199	488 332	894 531

could potentially last 340 years at the current rate of consumption assuming no new coal is found.

4.5 Projections

The usable energy in 186.9 trillion cubic meters of natural gas, the current global proven reserve, is about 6.9×10^3 EJ. Global petroleum reserves are at 1697.6×10^9 barrels of oil which is about 1.04×10^4 EJ. The energy equivalent of the world proven coal reserve is about 2.6×10^4 EJ. This represents some 4.3×10^4 EJ in total, and at the current global energy consumption rate of 552 EJ yr^{-1} we could depend on fossil fuels for another 78 years or so. But there are many reasons this is an unlikely scenario, as are the figures of 48 years until petroleum runs out, 25 years for natural gas and 340 years for coal given previously. As we have seen, new technology often results in sharp increases in the proven reserve, as happened with hydraulic fracturing and oil and gas reserves. It is also the case that when demand is high the price increases which drives conservation and innovation. Projections of future energy use are often grossly incorrect, as pointed out by Vaclav Smil (2005). As he shows in his book, various think-tanks and experts between 1960 and 1980 made projections of future US energy use for the year 2000 which ranged between a low of 110 EJ and highs above 190 EJ, with the majority projecting more than 150 EJ would be needed. The actual use in 2000 was about 100 EJ, so all of these projections were wrong and some spectacularly so. We should be very cautious about projecting how much energy we will need in the future and how much of any particular resources remains untapped.

It is easy to construct arguments and even mathematical equations that show the production of particular fossil fuel ‘peaking’ at some particular time in the relatively close future but it should be kept in mind that these are projections. No matter how sophisticated the math, they are not laws of physics but rather estimates of what might happen given what we know today (Hafemeister 2007). Arguably the most widely used tool for predicting the depletion of a finite resource (or even a renewable resource which is being used faster than it is replaced) is the Verhulst or logistic equation[5]. This method was developed following estimations of petroleum resources in the lower 48 states of the US made by M King Hubbert in 1956 (Hubbert 1957). He predicted a peak in oil production in the early 1970s for the continental US and was famously accurate, but many things can and have changed. Because of hydraulic fracturing, oil production in the continental US has recently risen to be as high as the 1970 peak, as shown in figure 4.8 (Marder *et al* 2016).

Calculations using the logistic equation (shown below) made in the past few decades show a peak year for global oil production in the current decade, and in fact global *conventional* oil extraction, not including unconventional or tight oil, did peak around 2010 (IEA 2014). But the amount of petroleum available on the global market has continued to increase because fracking has made previously inaccessible

---

[5] This is not the only model for predicting the lifetime of a fossil fuel source. For example see Shafiee and Topal (2009).

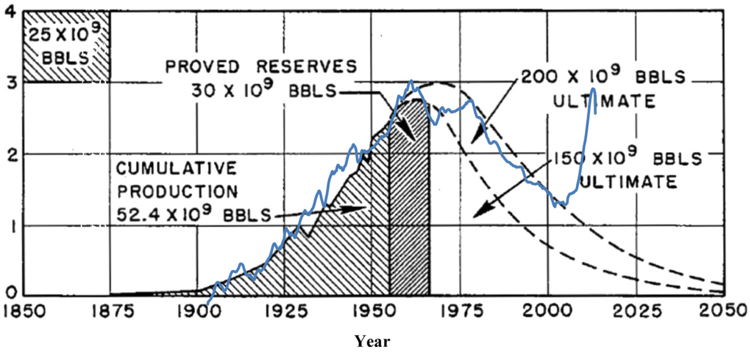


Figure 4.8. Hubbert's famous prediction for US oil production compared to the actual production (10 year running average) in billions of barrels per year (BBLs). US oil production did peak in 1972 as he predicted. The slight increase in the late 1970s and early 1980s was due to new Alaska oil. The sharp upturn in production after 2006 was due to fracking.

oil and natural gas available, particularly in the US. Other factors such as a switch away from petroleum in favor of natural gas as a primary energy source and better transportation fuel economy have decreased the importance and demand for oil globally. In the Paris Agreement of 12 December 2015, 195 countries committed to holding global warming to less than 2 °C which will require a reduction in carbon dioxide emissions, and this has already had an effect on global fossil fuel markets, reducing demand for coal and increasing demand and use of natural gas[6].

It is true, however, that fossil fuels are a non-renewable resource; no new petroleum, natural gas or coal is being formed. In the 1950s and 1960s the amount of new oil discovered each year was in the range of hundreds of billions of barrels. This has dropped to less than 20 billion barrels of new oil discovered per year, as shown in figure 4.9, which should be compared to the annual consumption of seven billion barrels of oil for the US alone. The world is not going to run out of oil or any other fossil fuel in the immediate future, but fossil fuels are not an infinite energy resource, even with new technological developments. The additional and possibly more important question of fossil fuel emissions in the form of carbon dioxide and its effect on the climate are discussed in chapter 9.

4.5.1 Other fossil fuels?

The prospect of finding other kinds of fossil fuels is, of course possible and in fact there has been an ongoing discussion for many years of *clathrate hydrates* which consists of methane trapped in ice. Clathrate is a generic term for a crystalline structure which contains trapped methane or other gases. They can only exist under high pressure and low temperature and in nature methane clathrate hydrates

---

[6] The agreement went into effect in November 2016 after being ratified by a majority of the participating countries. At the present time the current US administration is seeking to roll back this commitment.

are found on the deep sea floor, in arctic permafrost and ocean sediments. These deposits have long been known to exist and are often found when drilling for oil or natural gas in the deep sea floor or in arctic regions. Initial estimates of how much of this resource exists were as high as 3×10^{18} m^3 of methane, which would be on the order of a million times more energy than the estimated natural reserves cited in figure 4.7. A better understanding of how clathrates form and further surveys have revised that estimate downward by a factor of at least 1000. Still, methane clathrate could be a large potential energy source; more than current world fossil fuel proven reserves if it exists at the admittedly speculative estimates often cited. However because this resource is not found in concentrated reservoirs as oil, coal and natural gas are, and because it only forms in deep ocean or under frozen sediments, it is not clear how or even if it could be captured for human use (Williams *et al* 2005).

As with all fossil fuels, burning methane from clathrates will introduce carbon dioxide and other pollutants into the atmosphere. Connected with the concern about carbon dioxide if clathrates become an energy source is the fact that if the Earth continues to warm at the present rate there is the potential for enormous amounts of methane to be released into the atmosphere from melting clathrates due to thawing of permafrost and warming of the deep ocean.

4.5.2 Hubbert's calculation

The following is an explanation (following Hafemeister (2007)) of the method developed by Hubbert and others for predicting the peak in production of a finite resource or renewable resource being consumed faster than it is renewed. We start with the assumption that the discovery of oil increases exponentially from zero but then levels off (curve Q_D in figure 4.10) at some finite value, Q_∞ after which no more reserves are found. The cumulative production curve, Q_P, has a similar shape but lags behind the discovery curve by a few years as markets expand until new reserves are no longer being found. The amount remaining in the reserve as a function of time would be $Q_R = Q_D - Q_P$, which will drop to zero once all the discovered oil has been produced. The curve Q_R has been modeled as a Gaussian or normal

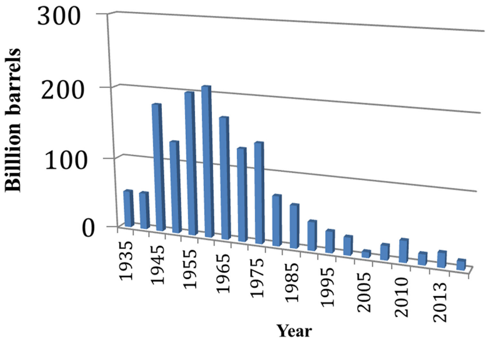


Figure 4.9. Global discoveries of conventional oil by year (Deffeyes 2001, Hafemeister 2007, EIA 2015 and EIA 2016a).

distribution curve, which is a reasonable guess. The model can be made more sophisticated as shown below. The area under the curve Q_R is the ultimate recoverable resource Q_∞—the total oil which will ever be recovered from the reservoir.

For exponential increase of a quantity, Q, $Q(t) = Q_0 e^{\lambda t}$. Here $Q(t)$ is the cumulative production at time t, Q_0 is the known cumulative production at some initial time which we will call $t = 0$ and λ is the exponential growth rate. The rate of production, $\frac{dQ}{dt}$ for exponential increase is thus given by

$$\frac{dQ}{dt} = \lambda Q_0 e^{\lambda t} = \lambda Q. \tag{4.2}$$

The rate of resource production can only increase exponentially initially. Once a significant proportion of the resource has been extracted the rate would be proportional to the fraction remaining, $1 - Q(t)/Q_\infty$, where Q_∞ is the total resource available assuming the cost of extraction and extraction techniques remain constant. We can therefore replace equation (4.2) with

$$\frac{dQ}{dt} = \lambda Q\left(1 - \frac{Q}{Q_\infty}\right), \tag{4.3}$$

which is known as the Verhulst or logistic equation. Equation (4.3) can be solved to give

$$Q(t) = \frac{Q_\infty}{1 + a e^{-\lambda t}}, \tag{4.4}$$

where $a = (Q_\infty - Q_0)/Q_0$ and Q_0 is the cumulative production at some initial year when $t = 0$. This solution has the same shape as either the discovery or production curve in figure 4.10 and is called the logistic curve. The derivative of $Q(t)$ gives the rate of production,

$$P = \frac{dQ}{dt} = \frac{Q_\infty a \lambda e^{-\lambda t}}{(1 + a e^{-\lambda t})^2}, \tag{4.5}$$

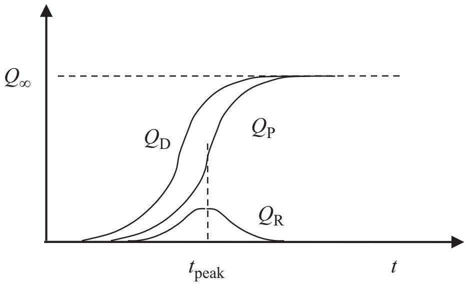


Figure 4.10. Theoretical resource discovery, Q_D production, Q_P and amount remaining, $Q_R = Q_D - Q_P$. The time of peak production, t_{peak}, is shown. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

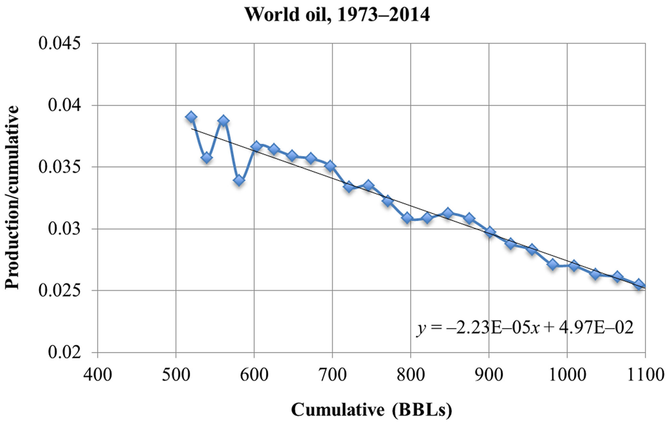


Figure 4.11. Global annual crude oil production divided by cumulative production versus cumulative production in giga-barrels (Gb = 10^9 barrels). Data from EIA (2016b).

which has a shape similar to the curve Q_R in figure 4.10 since $\lim_{t\to 0} \frac{Q_D - Q_P}{t} = \frac{dQ}{dt}$. The time of peak production, t_{peak}, can be found by setting the derivative of P with respect to time equal to zero and solving for t to obtain

$$t_{\text{peak}} = \frac{1}{\lambda} \ln a. \tag{4.6}$$

Initially Hubbert used the known cumulative production, Q_0 at some initial time and the proven reserves known at that time to hand fit either equations (4.4) or (4.5) to obtain an estimate of the ultimate resource available, Q_∞, the year of peak production, t_{peak} and the production during the year of peak production, $P(t_{\text{peak}})$. Trying to fit equation (4.4) to available data is difficult and can result in significant error, however, especially if cumulative production is not accurately known.

It is generally easier to match data to a straight line and this can be done with the Verhulst equation, by replacing the rate of production in equation (4.3) with $P = \frac{dQ}{dt}$ to obtain

$$P/Q = \lambda - \frac{\lambda Q}{Q_\infty}. \tag{4.7}$$

Plotting P/Q versus Q will give a straight line with slope $-\lambda/Q_\infty$ and the x intercept equals Q_∞. The only data needed to make this plot are the annual production, P, and the cumulative production, Q, over time. The total oil which can be extracted, Q_∞ and the production rate, λ, are produced from a graph of the data. Equation (4.6) gives the year of peak production where a is found by inverting equation (4.4) using data from a known year.

As an example, data of oil production from 1973 to 2014 from the EIA can be used to predict global peak oil (from conventional sources only). Multiplying the historical daily production by 365 days (and scaling to giga-barrels) gives the historical annual production, P. The cumulative production, Q, is found by starting with an estimate of the cumulative production up to 1973 and then adding each year's production to the sum of the previous years. A plot of P/Q versus Q using 500 Gb (Gb = 10^9 barrels) as the cumulative production in 1973 is shown in figure 4.11.

A straight line fit of the data (calculated with the linear fit function in Excel) is $y = 2.23 \times 10^{-5}x + 0.0497$. The global rate of oil production for this period of time, according to the model, is approximately $\lambda = 0.0497 \approx 5\%$. Solving the equation for $y = 0$ gives the x-intercept which is the total recoverable world supply of conventional oil: $Q_\infty = 2490$ Gb.

The peak year can be found by first solving equation (4.4) for a to obtain

$$a = \left(\frac{Q_\infty}{Q(t)} - 1\right)e^{\lambda t}. \tag{4.8}$$

Picking 1980 as year $t = 0$ and using $521.3\,\text{Gb} = Q_0$ from the data for the cumulative production in that year gives $a = (\frac{Q_\infty}{Q(0)} - 1)e^{\lambda 0} = 3.8$. Equation (4.6) then gives $t_{\text{peak}} = \frac{1}{\lambda}\ln a = 26.7$ years. So the predicted peak year of world oil production is $1980 + 26.7 \approx 2006$. According to IEA data, global conventional oil production did peak sometime around 2010[7], but because of hydraulic fracturing total production of conventional plus tight shale oil has continued to increase.

Based on current proven reserve figures and a logistic calculation, peak production of natural gas including gas from fracking would occur in 2025, peak petroleum including oil from fracking would be in 2033, a peak year for coal occurs in 2060 and a peak year for uranium in 2113. These forecasts are probably more useful than simply dividing the reserve by the rate used, but they are predictions, not inevitability. It is interesting to note that the logistic equation gives a single peak if the parameters of the model do not change but the same equation can give the curve shown in figure 4.8 for the real oil production in the US if the input parameters are adjusted differently (Hafemeister 2007).

4.6 Energy return on energy invested

Another way to think about the usefulness of energy resources is *energy return on energy invested* (EROEI), which is the amount of energy required to extract a unit of energy, refine it into a useable form and deliver it to market. There have been many estimates of this figure for different resources and in 2013 Hall *et al* performed a meta-analysis of available research on the EROEI, some of the results of which are shown in table 4.3. Clearly a resource that returned less energy than it takes to extract would not be very useful as an energy source, although there might be reasons for extraction if there is no alternative for that fuel. Currently jet fuel from petroleum is the only way to store sufficient energy in a fuel for long distance

---

[7] The exact year is difficult to determine because the peak in the data is relatively flat.

Table 4.3. EROEI for several energy resources[a]

Energy source	Approximate EROEI (Hall *et al* 2014)
US oil and gas production 1930s	100:1
US oil and gas production 1970s	25:1
US oil and gas production 2007	10:1
Global oil and gas production 1995	30:1
Global oil and gas production 2006	18:1
Tar sands	4:1
Shale oil	7:1
Global coal production	46:1
US coal production 1950s	80:1
US coal production 1980s	30:1
US coal production 1990s	80:1
Nuclear	14:1
Hydroelectric	84:1
Ethanol from biomass	5:1
Ethanol from corn	0.8–1.6:1
Ethanol from sugar cane	0.8–10:1
Diesel from biomass	2:1
Wind	18–20:1
PV solar	10:1
Geothermal	9:1

[a] Renewable energy resources also have much lower life-cycle emissions (total emissions in building and operating a power plant using these fuels) than fossil fuels. See (IPCC 2012) and chapter 6.

commercial jet planes, so it might make sense to extract petroleum for this purpose, even if it takes more energy to extract it than there is in the petroleum extracted.

As can be seen from the figures in table 4.3, the EROEI for a given fuel is not a fixed number. Initially in the US petroleum extraction only required drilling a well and the EROEI was 100 to 1. Secondary, tertiary and hydraulic fracturing methods require more energy and the EROEI for US domestic oil has fallen to approximately 10:1. We can expect the EROEI for tar sands, wind and PVs to increase over time as new technology makes these extraction methods more efficient. The low EROEI for biomass (ethanol from corn in particular) reflect basic limitations in the biochemistry of growing plants. Evolution has developed photosynthesis to be very efficient at capturing light but plants use much of this energy for growth and only some of it for storing energy as oils or high energy content biomass. Sugar cane produces about 10 times more fuel energy than corn per square kilometer and Brazil has been successful in supplying a large amount of its transportation needs from sugar cane as a result. Algae appears to be slightly better than many plants in turning sunlight into fuel, but theoretical maximum efficiency for photosynthesis appears to be less than 8% (Zhu *et al* 2008).

The reversal in the decline in EROEI for coal in the US between the 1980s and 1990s probably reflects the move from deep mining to surface mining, for example in the form of mountain top removal in the Appalachian mountains in the US. The EROEI for coal appears to have peaked around 2000 and is declining again as the ratio of high quality coal (anthracite) extracted to coal with less energy content decreases. Hydroelectric energy has the highest EROEI of all the renewables but it is still not as high as petroleum was when petroleum first came on the market as an energy source. It seems likely the other renewable energy sources will have rising EROEI, since, unlike petroleum extraction techniques, new renewable technology should make renewables more energy efficient.

Projects

1. Find the most recent BP data on energy (www.bp.com/). Compare some subset of these figures with similar data from the EIA (www.eia.gov/) and with the same data from the IEA (www.iea.org/) if available. How do they compare—do they agree? Why not?
2. Write a report on other uses of petroleum besides for fuel.
3. Write a report on methods used for searching for petroleum, natural gas and coal resources. Describe the physics involved.
4. Write a report on the effects of global political changes on the price of oil. Start with the events mentioned in the text and expand and elaborate on those.
5. Make a more fine-grained version of table 4.1, showing equivalent figures for several other countries including developed and developing countries. Write a description of how they are different from the figures in the table.
6. Break down the columns in table 4.1 to show typical entries into each column. (In other words, what do the various countries use electricity for? Is the same in each country? What is different about how developing countries use energy for transportation compared to developed countries? How are manufacturing processes different between countries?) Discuss your findings.
7. Write a report explaining why the future projections for energy use for the US (given in Smil 2005) turned out to be so far off. Who made these predictions? With what assumptions did they start? What was their reasoning? Is this likely to occur today?
8. Find examples of how the logistic equation has been applied to other resources. Describe how it might be used to predict 'peak wood production' or some other renewable resource which is being harvested faster than it is being renewed.
9. Verify the calculations of peak natural gas, oil, coal and uranium found in the text using BP data (www.bp.com/) or data from the IEA (www.iea.org) and the logistic equation.
10. Write a report on EROEI and how it is calculated. Include any criticisms and alternatives you find. Verify the figures given in table 4.3. What assumptions were made in these estimates?

References

BP 2016 *Statistical Review of World Energy* www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html (Accessed: 7 August 2016)

Deffeyes K S 2001 *Hubbert's Peak: The Impending World Oil Shortage* (Cambridge, MA: Princeton University Press)

The Economist 2015 Made in China? March 12

EIA (US Energy Information Administration) 2012 Annual Energy Review (Archives) https://www.eia.gov/totalenergy/data/annual/index.php (Accessed: 1 April 2017)

EIA (US Energy Information Administration) 2015 Financial Review of the Global Oil and Natural Gas Industry 2015 www.eia.gov/finance/review/pdf/financial_2015.pdf (Accessed: 7 August 2016)

EIA (US Energy Information Administration) 2016a Annual Energy Outlook www.eia.gov/forecasts/aeo/index.cfm and www.eia.gov/forecasts/ieo/ieo_tables.cfm (Accessed: 12 August 2016)

EIA (US Energy Information Administration) 2016b International Statistics Data www.eia.gov/totalenergy/data/monthly/pdf/sec11_5.pdf (Accessed: 7 January 2017)

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Gray M, Xu Z and Masliyah J 2004 Physics in the oil sands of Alberta *Physics Today* **57** 7 p 47

Hafemeister D 2007 *Physics of Societal Issues* (New York: Springer)

Hall C A S, Lambert J G and Balogh S B 2014 EROI of different fuels and the implications for society *Energy Policy* **64** 141–52

Hubbert M K 1957 Nuclear energy and the fossil fuels *Shell Development Publication* 95 (presentation to the Spring Meeting of the Southern District Division of Production, American Petroleum Institute, March 7–9)

IEA (International Energy Agency) 2014 *World Energy Outlook 2014* http://www.iea.org/publications/freepublications/publication/WEO2014.pdf (Accessed: 2 April 2017)

IEA (International Energy Agency) 2016 www.iea.org/bookshop/728-Renewables_Information_2016 (Accessed: 10 September 2016)

IPCC (International Panel on Climate Change) 2012 *Renewable energy storage and climate change mitigation* www.ipcc.ch/pdf/special-reports/srren/SRREN_Full_Report.pdf (Accessed: 6 January 2017)

Ma H 2014 Global energy and carbon intensity continue to decline world watch vital signs Dec 15 http://vitalsigns.worldwatch.org/vs-trend/global-energy-and-carbon-intensity-continue-decline (Accessed: 12 August 2016)

Marder M, Patzek T and Tinker S W 2016 Physics fracking fuel and the future *Physics Today* **69** 7 p 46

Shafiee S and Topal E 2009 When will fossil fuel reserves be diminished? *Energy Policy* **37** 181–9

Smil V 2005 *Energy at the Crossroads: Global Perspectives and Uncertainties* (Boston, MA: MIT Press)

Weingarten M S, Ge S J W, Godt J W, Bekins B A and Rubinstein J L 2015 High-rate injection is associated with the increase in US mid-continent seismicity *Science* **348** 1336–40

Williams T E, Millheim K and Liddell B 2005 Methane hydrate production from the alaska permafrost *department of energy* DE-FC26-01NT41331 March 2005

Zhu X, Long S and Ort D 2008 What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Curr. Opin. Biotechnol.* **19** 153

IOP Concise Physics

Physics and the Environment

Kyle Forinash III

Chapter 5

Nuclear energy

The existing 449 nuclear power plants globally supply about 5% of the world's primary energy and almost 11% of its electricity (IAEA 2016). France relies on nuclear power for 76% of its electricity and 12 other countries depend on nuclear reactors for more than 30% of their electricity. About 60 new nuclear power plants are under construction, mostly in Asia. Nuclear power plants typically remain online more than 90% of the time, having to shut down very infrequently for maintenance, compared to capacity factors for wind and solar of 40% and 20%, respectively[1]. As we will see in the next chapter (table 6.3) the projected levelized cost of electricity from new power plants is significantly lower for nuclear than coal with carbon sequestration and gas turbines, which makes it economically attractive. Nuclear is also cleaner than fossil fuels in regards to pollution as we discuss below (Markandya and Wilkinson 2007). All of these factors make nuclear power plants attractive as a base load supply of electricity. Other sources which are more expensive but can be brought online more quickly than nuclear can then be used to supplement this base load source. This chapter reviews how nuclear power works and how it can be put to use as a primary energy source.

5.1 Nuclear reactions

The number of protons in an atomic nucleus is called the *atomic number*, Z, and the number of neutrons is given by the neutron number, N. Adding these two together gives the number of nucleons in the nucleus which is called the *mass number*, A. This is also the approximate mass of the nucleus in *atomic mass units* since each nucleon

---

[1] Capacity factor is defined to be the ratio of the actual output of a power plant to its theoretical maximum output if it could run continuously over some period of time. For example PV solar cells have a theoretical output based on the efficiency of the cell and the number of hours of sunlight. They will not achieve this maximum output because of various operational issues such as weather, daily changes in Sun angle and maintenance.

doi:10.1088/978-1-6817-4493-3ch5

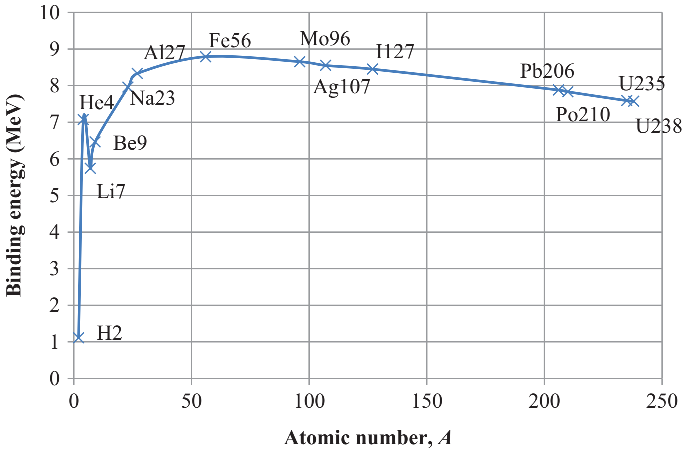


Figure 5.1. Binding energy per nucleon curve. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC. 1 MeV = 10^6 eV, where 1 eV = 1.6×10^{-19} J.

has a mass of one atomic mass unit[2]. Elements in the periodic table can be specified by their mass number, atomic number and a letter with the notation $^A_Z X$ where X denotes the chemical symbol of the element (and is redundant since the atomic number is given). The chemical behavior of an element is determined by the number and configuration of the electrons and a neutral atom has the same number of electrons as protons. Carbon $^{12}_{6}\mathrm{C}$ and $^{14}_{6}\mathrm{C}$ have the same chemical properties since they have the same number of protons (and therefore the same number and configuration of electrons) but different numbers of neutrons. Electrons are designated as e^- and neutrons as $^1_0\mathrm{n}$ in the following. Nuclear reactions also involve short lived positrons, e^+, which are in effect positively charged electrons. The neutrino, ν, and anti-neutrino, $\bar{\nu}$, are also created in nuclear reactions and were discovered by conservation of momentum considerations applied to nuclear decay processes. Neutrinos are nearly massless particles which do not interact very strongly with normal matter, making them very difficult to detect.

The strong nuclear force (generated by quark[3] interactions) is a short range force which is much stronger than the electrical repulsion of the protons and holds the nucleus together. Neutrons in the nucleus also interact with each other and with protons via the strong force but do not have electrical charge. For heavier elements in the periodic table the ratio of neutrons to protons is larger in order to provide more nuclear force needed to hold the nucleus together against the electrical repulsion of the protons. Very large nuclei are not stable for this reason. Nuclear binding energy per nucleon is defined to be the energy needed to separate all the

---

[2] Defined to be one twelfth the mass of a carbon atom: 1.66×10^{-27} kg.

[3] Neutrons and protons each consist of three smaller particles called quarks which are held together by an exchange of massless particles called gluons. Much of nuclear theory was developed before the discovery of quarks with the result that quark theory is not required to explain most nuclear reactions. The strong nuclear force quantifies the interaction between quarks inside a nucleon with the quarks in another nucleon.

Table 5.1. The four types of radioactive decay. 1 eV = 1.602×10^{-19} J.

Process	Example	Approximate energy released
Alpha decay	${}^{238}_{92}\mathrm{U} \rightarrow {}^{234}_{90}\mathrm{Th} + {}^{4}_{2}\mathrm{He}$	4 MeV
Beta decay	${}^{14}_{6}\mathrm{C} \rightarrow {}^{14}_{7}\mathrm{N} + \mathrm{e}^{-} + \bar{\nu}$	9 MeV
	${}^{12}_{7}\mathrm{N} \rightarrow {}^{12}_{6}\mathrm{C} + \mathrm{e}^{+} + \nu$ (requires energy)	
Gamma decay	${}^{A}_{Z}X^{*} \rightarrow {}^{A}_{Z}X + \gamma$	4 MeV
Fission	${}^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{141}_{56}\mathrm{Ba} + {}^{92}_{36}\mathrm{Kr} + 3{}^{1}_{0}\mathrm{n}$	200 MeV
	${}^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{140}_{54}\mathrm{Xe} + {}^{94}_{38}\mathrm{Sr} + 2{}^{1}_{0}\mathrm{n}$	

protons and neutrons in a nucleus from each other divided by the number of nucleons present. Figure 5.1 shows a plot of binding energy per nucleon versus the mass number. Iron has the highest nuclear bonding energy per nucleon and thus has the most stable nucleus of any element.

All elements heavier than iron have lower binding energies per nucleon which means the nuclei are less stable. For these heavier elements there is a potential to undergo a nuclear decay process, giving off energy in order to move to a more stable nuclear configuration. These decay processes convert tiny amounts of mass into energy according to the well-known equation, $E = mc^2$ where $c = 2.999 \times 10^{8}\ \mathrm{m\,s^{-1}}$ is the speed of light. Due to the large value of c, only a small amount of mass is needed to provide large amounts of energy[4]. The four types of decay processes are shown in table 5.1 with the approximate energy released for that process. It should be noted that chemical reactions are on the order of a few electron volts per atom whereas nuclear reactions are in the range of 10^{6} eV (MeV).

Two protons and two neutrons, which form a helium nucleus, are given off in *alpha decay*. Electromagnetic waves with frequencies above 10^{20} Hz are called *gamma radiation* and occur when the nuclei in an excited nucleus (designated by $X*$) relax to a lower energy levels. Radioactive carbon dating is based on *beta decay* which is the emission of an electron and a neutrino. The same process can occur with the emission of a positron which, because of conservation of charge, requires that the atomic number decrease by one without changing the mass number. Nuclei can also split into two parts in a process called *fission*. This can occur spontaneously or be prompted by a collision with a neutron as shown by the two examples of ${}^{235}_{92}\mathrm{U}$ decay in table 5.1. Other fission processes are possible. Part of the problem of nuclear waste comes from the fact that the splitting which occurs during fission can lead to many different daughter nuclei, some of which are chemically toxic or have unstable nuclei and will give off one of the other types of radiation in the table.

Some isotopes are more likely to take part in a decay process than others and the exact time when an individual nucleus will spontaneously decay is not predictable by

---

[4] This is why the word 'approximate' was used above when equating mass number, A, with the mass in atomic mass units. Bound nucleons have a slightly different mass than when they are not in a nucleus which is accounted for by the binding energy.

the laws of physics. However a large number of identical nuclei will collectively obey an exponential decay law according to the formula

$$R = \lambda N_0 e^{-\lambda t}, \tag{5.1}$$

where R is the activity, measured in becquerels (Bq): 1 Bq = 1 emission s^{-1}. Here N_0 is the orginal number of radioactive atoms and the original activity is $R_0 = \lambda N_0$, where λ is the decay constant in units of 1/s. The half-life is defined to be the time needed for half of a radioactive sample to decay and is related to the decay constant by

$$t_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda}. \tag{5.2}$$

Half-lives of various decay processes can vary from fractions of a second to millions of years, a property that makes nuclear waste disposal particularly difficult.

Because radioactive decay proceeds at a known rate for a given isotope it can be used to date rocks and artifacts. The first application of *radiometric dating* was done with carbon which has two chemically identical isotopes. The heavier isotope, ${}^{14}_{6}C$, undergoes beta decay with a half-life of $t_{1/2} = 5730$ yr. This isotope is created at a more or less constant rate in the upper atmosphere from nitrogen being struck by cosmic rays from space. There is, therefore, a constant ratio of ${}^{14}_{6}C$ to ${}^{12}_{6}C$ in atmospheric CO_2 of approximately 1.3×10^{-12}, which will also be the ratio of ${}^{14}_{6}C$ to ${}^{12}_{6}C$ found in living organisms since all organisms continue to take in CO_2 as they respire, maintaining a constant ratio. Once an organism expires, however, the ${}^{14}_{6}C$ begins to decay and the ratio changes. Since this occurs at a known rate, the approximate date of death can be determined by comparing the ratio of carbon isotopes in the dead organism to the atmospheric constant ratio of 1.3×10^{-12}.

As an example, suppose an organic artifact such as a wooden handle or piece of cotton cloth is found in an archeological site and the age of the site is to be determined. If the ${}^{14}_{6}C$ activity is measured to be 7.0 Bq in a sample which contains 50.0 g of carbon we can calculate the approximate time the wood or cotton was harvested. The decay constant for ${}^{14}_{6}C$ is $\lambda = \frac{0.693}{t_{1/2}} = 3.84 \times 10^{-12}\ s^{-1}$. Using the known atmospheric ratio the sample had $50.0\ g \times 1.3 \times 10^{-12} = 6.5 \times 10^{-11}$ g of ${}^{14}_{6}C$ when it was alive. A mole of ${}^{14}_{6}C$ weighs 14 g so the original number of ${}^{14}_{6}C$ molecules is $N_0 = 6.5 \times 10^{-11}\ g \times 6.02 \times 10^{23}$ atoms/14 g $= 2.8 \times 10^{12}$ atoms. Using equation (5.1) we have $7.0\ Bq = 2.8 \times 10^{12} \times 3.84 \times 10^{-12}\ s^{-1}\, e^{-(3.84\times10^{-12}\ s^{-1})t}$. Solving for t gives $t = 1.1 \times 10^{11}$ s = 3529 yr, so the discovered material was harvested some 3500 years ago.

The half-life of ${}^{14}_{6}C$ is 5730 yr which is not suitable for material older than about 50 000 years, but many other nuclear decay processes with longer half-lives can be used to date older samples. Uranium decays into lead via a combination of alpha and beta decays and this occurs at a known rate. The ratio of uranium to lead in a rock sample will be proportional to the age when the rock was formed and can be used to date the sample. Similar processes which have been used for dating purposes include an isotope of samarium, which decays into neodymium, potassium which decays into argon, uranium decay into thorium and a rubidium isotope which will decay into strontium. Each of these decay processes has a different half-life, providing tools for examining the age of material which formed at different times.

5.2 Nuclear reactors

How can the reactions shown in table 5.1 be harnessed as energy supplies? In the fission process we notice that if an atom of $^{235}_{92}\mathrm{U}$ is struck by a neutron it will fission into two parts, giving off two or three new neutrons and 200 MeV of energy. Other $^{235}_{92}\mathrm{U}$ atoms can be arranged so that the neutrons from the first reaction will result in two or three more fissions or 600 MeV of energy and eight or nine more neutrons. This process of neutrons being emitted from a fission reaction which subsequently causes more decays and more emitted neutrons is called a *chain reaction* and is the source of energy in nuclear reactors and atomic bombs. If the first fission reaction in table 5.1 proceeds indefinitely the energy given off grows as 3^n where n is the number of decays. Obviously for energy production we do not want the process to continue indefinitely, which is called a *supercritical* chain reaction. A reaction where the number of neutrons remains the same at each stage because some of them do not collide with another nucleus or do not have the required energy is said to be *critical*. If fewer neutrons occur at each stage the reaction is *subcritical* and the reaction eventually dies out. One way this can happen is if the amount of $^{235}_{92}\mathrm{U}$ is insufficient, in other words a *critical mass* of $^{235}_{92}\mathrm{U}$ is needed so that each new neutron given off finds a nucleus to collide with. It should be noted that other isotopes of uranium do not fission. There must be a critical mass of the isotope $^{235}_{92}\mathrm{U}$ present for a chain reaction to occur.

The neutrons emitted from the fission reactions are generally too fast to be effective in stimulating further emissions and will bounce off the nucleus rather than cause a decay. In most reactors the neutrons are slowed down by adding a *moderator* such as water, graphite or heavy water (water where one hydrogen is an isotope with an extra neutron). Once the fast neutrons have collided with the moderator atoms they are slower and are called *thermal neutrons* and the reactor is referred to as a thermal nuclear reactor. The few reactors which do not have moderators are denoted as fast reactors because they use the faster neutrons. These reactors generally can be smaller and have less radioactive waste, but are more difficult to build and require a higher percentage of $^{235}_{92}\mathrm{U}$ (or $^{241}_{94}\mathrm{Pu}$ which undergoes a similar fission process) in order to supply a larger neutron flux to make up for the fact that fewer fast neutrons cause fission.

In nuclear reactors the $^{235}_{92}\mathrm{U}$ is arranged in long thin fuel rods so that the chain reaction chain reaction cannot continue to be supercritical past a given point. Because the concentration of fissionable uranium is relatively low, most reactors also require a surrounding material called a *reflector* which returns some of the neutrons to the uranium fuel in order to maintain the chain reaction at the critical level. *Control rods*, which are made from substances that absorb neutrons such as silver–

indium–cadmium alloys, are placed in the space between the fuel rods and can be raised or lowered to control the number of neutrons passing from one fuel rod to the next, thus controlling the rate of the chain reaction[5].

More than 65% of the currently functioning nuclear reactor are pressurized light water reactors (PWR). The energy emitted by the radioactive fission of uranium in these reactors is captured as thermal energy by ordinary water under high pressure (>150 atmospheres) and used to form steam. The water acts as both moderator, to slow the emitted neutrons down to the right speed for further decay, and heat transfer medium[6]. Fuel in this kind of reactor is in the form of uranium pellets, about the size of a dime with a concentration of $^{235}_{92}\mathrm{U}$ of about 4%. The pellets are arranged in a fuel rod about one centimeter in diameter and three to four meters in length and fuel rods are arranged in bundles of 250 rods. A typical reactor core has around 200 fuel rod assemblies. Control rods that absorb neutrons, are interspersed between the fuel rods.

Another 23% of currently operating reactors are boiling water reactors which work in basically the same way as a PWR but the water is not under pressure and forms steam directly. Other reactor designs in current use include the Canadian deuterium uranium reactor (CANDU), pressurized heavy water reactor, RMBK (a channelized large power boiling reactor—Chernobyl type) and advanced gas cooled reactor. Reactors have been designed with no moderators (fast reactors) and with various types of coolants such as water, pressurized water, inert gas, or liquid metals such as sodium. Plutonium and un-enriched uranium can be used as fuel (the CANDU reactor) and thorium has been considered as a potential fuel. Details of other reactor designs can be found in (Lamarsh and Baratta 2001).

Once steam has been formed a nuclear power plant works exactly the same as a coal or natural gas fired plant; steam turns a turbine connected to an electric generator to make electricity. As in fossil fuel power plants, the water must first be cooled before being returned to the reactor for another cycle. The efficiency of the steam turbine–electrical generation process is limited by the second law of thermodynamics just as any other thermal process is.

Reactor designs have been categorized into four generations with generation II representing most existing reactors. There have been incremental improvements over time in these designs, which are termed generation III reactors. Generation IV reactors are new designs which have yet to be implemented on a commercial scale. Generation IV reactors use less fuel, are smaller, less expensive, have fewer moving parts and are planned to be inherently safe, meaning that they will automatically shut down under their own power and cool using natural convection in the event of operator error, mechanical failure or failure of external and backup electric power. An example of this design is the gas cooled pebble bed reactor. The fuel for this reactor is in the form of thousands of billiard ball sized spheres made of a carbon

---

[5] In an emergency shutdown of a reactor (called a scram) the rods are allowed to drop into the reactor under the force of gravity to stop the chain reaction.

[6] The water transfers heat but not radioactivity from inside the reactor core to make steam outside the core which turns steam turbines.

ceramic compound which encase mustard seed sized grains of uranium or other fissionable material. The reactor never needs to be shut down for refueling since fuel balls can continuously be added at the top and spent fuel removed from the bottom. The balls are coated with a tough, high temperature resistant chemically inert shell and can be disposed of without further processing. The casing makes it impractical to use the fuel in an improvised nuclear weapon. The heat transfer medium is helium, which means that the risk of steam explosion is eliminated and the reactor can operate at higher temperatures, making the second law thermal efficiency higher. Two experimental reactors with this design have been constructed, one in Germany (now decommissioned), the other in China (the HTR-10).

Reactors can be made with other radioactive isotopes such as ${}^{233}_{92}\mathrm{U}$ and ${}^{239}_{94}\mathrm{Pu}$ but these elements do not occur in nature because of their short half lives. They can be created, however in a so-called *breeder reactor* using the following process which occurs in a uranium reactor:

$${}^{238}_{92}\mathrm{U} + \mathrm{n} \rightarrow {}^{239}_{92}\mathrm{U} \rightarrow {}^{239}_{93}\mathrm{Np} + \mathrm{e}^{-} + \bar{\nu} \rightarrow {}^{239}_{94}\mathrm{Pu} + \mathrm{e}^{-} + \bar{\nu}. \tag{5.3}$$

The isotope ${}^{239}_{94}\mathrm{Pu}$ will undergo a fission in a manner similar to ${}^{235}_{92}\mathrm{U}$. Notice that the process uses non-fissionable ${}^{238}_{92}\mathrm{U}$, which makes up more than 95% of the uranium fuel rod and in fact this reaction provides most of the energy in a typical uranium reactor. The amount of plutonium being created can be enhanced by placing a layer of ${}^{238}_{92}\mathrm{U}$ around the core of a reactor, in front of or in place of the reflector, so that neutrons escaping from the core can be used to trasmute this material into fissionable fuel. All reactors are breeder reactors to some extent since fissionable ${}^{239}_{94}\mathrm{Pu}$ is being produced via the reaction in equation (5.3), but only those reactors designed so the produced fuel can be removed for use in other reactors are referred to as breeder reactors.

Fission is not the only nuclear reaction shown in table 5.1 which can be used as an energy supply. For example a small alpha source such as ${}^{238}_{94}\mathrm{Pu}$ can be used to generate heat which can be turned into electricity using a thermocouple. Most of the deep space probes which traveled to distances too far from the Sun to make photocells a viable energy source have used alpha or beta sources for energy. Examples include the Pioneer, Voyager and Cassini probes to Jupiter, Saturn, Uranus, Neptune and Pluto, respectively. Unlike for fission, the amount of shielding required for alpha decay is small making it an idea lightweight long term energy source. It is difficult, however, to scale this source up to the size of existing commercial power plants.

5.3 The refining process

World reserves of economically recoverable uranium are estimated to be 5.4 million tonnes with Australia having the largest deposits, approximately 30% of the world's resource. Uranium ore ranges from as high as 20% uranium in Canadian deposits to 0.01% in ore found in Namibia. The Earth's crust averages 2.8 parts per million (ppm) of uranium and ocean water is approximately 0.003 ppm uranium, but these sources are currently not counted as reserves because they are not considered economically viable at current uranium prices. Currently the world uses 66 000 tonnes of uranium per year (World Nuclear Association 2016).

The uranium in naturally occurring uranium deposits is typically about 1% $^{235}_{92}\mathrm{U}$ and 99% $^{238}_{92}\mathrm{U}$ and as we saw in the last section, it is the $^{235}_{92}\mathrm{U}$ which is useful for reactors. Although one reactor design, the CANDU reactor, has been built to use un-enriched uranium and other radioactive isotopes such as thorium have been proposed as reactor fuels, the majority of existing reactors use uranium enriched to 4% $^{235}_{92}\mathrm{U}$. Atomic bombs need to have more than 20% $^{235}_{92}\mathrm{U}$ to detonate and typically use uranium that is 80% $^{235}_{92}\mathrm{U}$. The fact that nuclear reactors use 4% enriched uranium means it is impossible for a reactor to explode like an atomic bomb, although other kinds of problems can occur[7].

The manufacturing process starts with uranium in the form of U_3O_8 (called yellowcake) which is separated from the material in which it is found by either a milling process or an *in situ* leaching process using acids. Because the chemical reactions for both isotopes of uranium are identical they must be separated by a physical process. Once the yellowcake has been separated from the ore in which it is found, it is chemically reacted with fluorine to make UF_6, a gas. Two methods are currently available to separate the two isotopes at this stage, gas diffusion and gas centrifuge.

Gas diffusion depends on the slight speed difference between the two isotopes in gaseous form at the same temperature. Recall that temperature is proportional to the average kinetic energy of a sample: $\langle \frac{1}{2}mv^2 \rangle = \frac{3}{2}kT$ where k is Boltzmann's constant and temperature is in kelvins (equation (2.4)). The lighter isotope, having a slightly higher average speed, will diffuse faster through a permeable pipe, separating the two isotopes. This process can be applied repeatedly until the desired concentration is achieved.

The gas centrifuge method also depends on the slight difference in mass between the two isotopes. Uranium hexafluoride gas is fed into a centrifuge and the heavier isotope will preferentially settle towards the outside of the spinning container. The lighter isotope can be removed from the region closer to the center of the centrifuge. The centrifuge method has replaced the gas diffusion method as it is more efficient in terms of the amount of energy needed to extract a given amount of enriched uranium.

Other methods of separation have been proposed, for example using lasers to preferentially push the lighter isotope away from the heavier one or ionizing the uranium metal and using electric fields to separate the two isotopes. Other techniques include the Becker nozzle in which the two isotopes in gaseous form (UF_6) flow over a curved surface so that the lighter isotope will follow a different circular path, much like the principle behind a mass spectrometer but using velocity differences instead of magnetic fields to separate the isotopes. These alternative

---

[7] For example in the case of the Chernobyl accident an intense graphite fire lofted radioactive material high into the atmosphere allowing it to spread over large parts of Europe. This reactor design is no longer being built.

methods are under development and the details are classified and/or proprietary. The technical difficulty in separating fissionable uranium from its more common isotope has made the threat of a terrorist group having an atomic device much less likely.

Once the uranium hexafluoride has been enriched to 4% it is chemically reacted to form a solid (UO_2), which is then manufactured into dime sized fuel pellets. The pellets then are stacked into fuel rods as previously described.

The $^{235}_{92}$U in a fuel rod gradually decays until the concentration decreases to less than 4%. Although only about 1% of the fuel has been used, a chain reaction cannot be sustained and in the US the fuel rods are then removed and put into temporary storage[8]. This process, referred to as an open fuel cycle, was adopted in the late 1970s, initially because it was feared that large scale recycling of fuel rods would become a national security risk if the technology became widely available. France and Japan, however, practice closed fuel cycles in which the uranium and plutonium from spent fuel rods is extracted using methods similar to the original enrichment process for uranium ore.

5.4 Radioactive waste

Radioactive waste is classified as *low level*, which includes material used in hospital treatments and research laboratories; *medium level* which includes nuclear reactor parts and some material used in industry; and *high level* which includes spent fuel rods and enriched fuels. Many proposals have been made for disposal of nuclear waste, including using existing reactors to transmute the material into other radioactive elements with shorter half-lives, recycling the material in a closed fuel cycle, and disposal either in ocean trenches or deep geological formations such as bore holes or salt domes or even into outer space or the Sun. The two methods currently used to dispose of most of the nuclear waste being generated is near-surface disposal and storage in deep, stable geological structures. A chemically stable solid compound in the form of rock or glass is formed by reacting the waste with other elements and the stabilized material is either stored on-site or buried in a dry rock formation such as a salt mine, in a zone where there is not likely to be much earthquake activity.

The majority of nuclear waste found in storage today comes from military defense projects but waste from commercial reactors is a growing problem. The daughter products from the two fission processes shown in table 5.1 are only a small sample of such fission products. Half-lives for these products vary substantially, ranging from one year for $^{106}_{44}$Ru to 2.1×10^5 yr for $^{99}_{43}$Tc, so that very long storage times are required. The fuel rod coatings and structural elements of the reactor become radioactive over time because the constant bombardment by fission products creates new isotopes. The half-lives of these new heavy nuclei range from 13 yr for $^{241}_{94}$Pu to 2.1×10^6 yr for $^{237}_{93}$Np. The nuclear bombardment process can also create elements which are chemically poisonous to humans and chemically reactive, which makes

---

[8] A solution for long term storage of radioactive material from nuclear power plants has yet to be agreed upon in the US.

Table 5.2. Old and new radiation units. The gray and the sievert are the currently accepted international units although all of these units appear in the literature.

Quantity	Unit(s)	Definition
Rate of decay	curie	1 Ci = 3.7×10^{10} decay s^{-1}
	becquerel	1 Bq = 1 decay s^{-1}
Energy intensity	roentgen	1 R = 87 erg g^{-1} = 0.0087 J kg^{-1}
Absorbed energy	rad	1 rad = 100 erg g^{-1} = 0.01 J kg^{-1}
	gray	1 Gy = 1 J kg^{-1} = 100 rad
Relative biological effectiveness	RBE	RBE = 1 for x-rays, γ-rays and β RBE = 5–20 for neutrons RBE = 20 for α and fission fragments
Effective absorbed dose	rem	rem = rad × RBE
	sievert	1 Sv = 1 J kg^{-1} = 100 rem

them hard to dispose of, even when they are not radioactive. A satisfactory solution for how to keep these dangerous byproducts of nuclear energy sequestered from humans for very long periods of time has not yet been found.

5.5 Radiation

Table 5.2 defines several terms relating to the measurement of radioactive decay and its effects. The *activity* of a substance (number of decays per second) was originally measured in *curies* but has been replaced by the *becquerel*. The energy released in a radioactive decay is large enough to ionize atoms leaving them charged. A *roentgen* is defined to be the amount of ionizing radiation that produces an electric charge of 3.33×10^{-10} C in 1 cm^3 of air under standard conditions. The older unit of *rad*, which is defined as the absorption of 0.01 J of energy per kilogram has been replaced by the *gray* which is equal to 100 rad. Different kinds of radiation with the same energy have different biological effects, which is given by the relative biological effectiveness (RBE). Multiplying the RBE times the absorbed radiation gives the effective absorbed dose measured in *sieverts* which has replaced the roentgen equivalent man (rem), an older unit. Unfortunately the literature about the effects of radiation contains a thorough mixture of old and new unit systems.

Radiation in the form of electromagnetic photons or high energy particles can be divided into two energy regimes. *Non-ionizing* photons do not have enough energy to disrupt normal biological processes whereas *ionizing radiation* does. As Einstein showed, each electromagnetic photon has an energy proportional to its frequency, $E = hf$ where h is Planck's constant. We do not have to worry that visible light, radio, cell phone or Wi-Fi signals will cause biological damage because the photons involved do not have sufficient energy to ionize molecules or atoms. On the other hand, x-ray and gamma photons have frequencies high enough so that the energy of each photon can do biological damage[9].

Alpha particles are easily blocked by as little as a piece of paper but if these particles do strike a living cell they are much more effective at causing damage than x-rays or electrons (β) due to their relatively large size. They are assigned an RBE of 20 for this reason. If the alpha source is inhaled or otherwise ingested it can be much more lethal than a beta or x-ray source of the same energy[10]. Neutrons have an intermediary RBE and x-rays, gamma and beta sources have low RBE. On the other hand gamma rays, x-rays and even neutrons require more shielding to prevent unwanted exposure. Beta particles, because they are charged, are blocked by a thin sheet of aluminum or other metal.

Radiation therapy makes use of the ionizing properties of radioactive decay. Fast growing tumors can be targeted by beaming ionizing radiation at them or by administering a compound containing a radioactive isotope which will chemically bind to the unique chemical structure of the tumor.

Naturally occurring radiation exposes the average person to 1.5 mSv to 3.5 mSv per year, although some locations, because of local deposits of radionuclides, expose the residents to levels as high as 130 mSv per year. North Americans receive an average radiation dose of about 3.6 mSv per year. A little less than 3 mSv of this exposure comes from natural causes such as cosmic rays and radon[11], a radioactive daughter product of uranium. The remaining exposure is from manmade sources such as medical x-rays. We cannot escape natural exposure to radiation. Potassium, calcium and carbon are essential elements for life and occur in our bodies and most of the foods we eat. Each has naturally occurring radioactive isotopes which are not chemically distinguishable from the more common isotope. For example, the potassium 40 content (a beta source) is 130 *becquerel* per kilogram for bananas, 125 Bq kg^{-1} for carrots and potatoes, 135 Bq kg^{-1} for lima beans, 207 Bq kg^{-1} for Brazil nuts and 14 Bq kg^{-1} for beer.

Table 5.3 gives exposure rates for several natural and manmade sources. Note that artificial exposure such as medical diagnostic tests, living at a high altitude (which increases cosmic ray exposure) and flying, are much larger sources of daily radiation exposure than natural background exposure.

The effects of high levels of radiation exposure are mostly known from the single doses received by victims of the Hiroshima and Nagasaki bombings during the Second World War and workers exposed during the Chernobyl disaster. Longer term exposure data have come from miners exposed to radon. Extremely high doses

---

[9] An analogy might be that a gamma photon is like a bullet where a single one can kill because of its high energy. Cell phone signals are more like tennis balls in that being struck by even a large number of them is not life threatening because each individual tennis ball is of low energy.

[10] Alexander Litvinenko, a former Russian security officer, was poisoned and killed with the polonium 210 isotope, an alpha emitter.

[11] Radon, ${}^{222}_{86}\mathrm{Rn}$, is a naturally occurring radioactive gas with a half-life of 3.8 days. In areas where ground concentrations are high, radon can seep into houses through cracks in basement walls or floors where it poses a lung cancer risk if the indoor concentration is sufficiently high (in fact it is the second leading cause of lung cancer after smoking in the US).

Table 5.3. Typical radiation exposure (Hafemeister 2007).

Source	Exposure (mSv/year)
One CT scan	10
Annual exposure, airline flight crew	2–9
Indoor radon	2.0
Cosmic radiation	0.31
Background radiation, sea level	0.28
Background radiation, Denver	0.81
Natural radiation in the body	0.39
Potassium 40 in the body	0.17
Single mammogram	0.3
Single dental x-ray	0.4
Single chest x-ray	0.1
US annual average x-ray exposure	0.53
One round trip transcontinental flight	0.05

Table 5.4. Exposure and excess cancer rates of Hiroshima and Nagasaki survivors (Hafemeister 2007). Expected cancers are the number of cancers expected for a similar population of the same size but with no excess radiation exposure.

Dose (Sv)	Victims	Expected cancers	Actual cancers	Excess cancers
<0.01	42 702	4267	4286	19
0.01–0.1	21 479	2191	2223	32
0.1–0.2	5307	574	599	25
0.2–0.5	5858	623	759	136
0.5–1.0	2882	289	418	129
1.0–2.0	1444	140	273	133
>2	300	23	55	32

of ionizing radioactivity cause changes in body chemistry which lead to death relatively quickly. A one time dose of 4–5 Sv kills 50% of the victims within a month, a dose greater than 6 Sv is lethal to 100% of the victims in the same time period. The cancer rates of Hiroshima and Nagasaki survivors are displayed in table 5.4 and a plot of dose versus cancer rate (excess cancers divided by number of victims) shows an approximate linear dose model for a one time large exposure with a slope of 0.45 cancers per sievert exposure (see figure 1.9).

Effects of smaller doses of radiation (less than 100 mSv) can be extrapolated from animal studies, uranium miners and populations with varying amounts of natural exposure. A single low dose of radiation above the natural background shows an increased lifetime cancer risk of 0.08 cancers per sievert exposure for men and 0.13/Sv for women, much lower than the 0.45/Sv cancer rate of nuclear bomb survivors (NRC 2006, Halperin 2007). This would indicate that low doses are less dangerous than large doses, but it is not clear if this holds for all types of cancers. Different tissues in the body have different susceptibility to radiation. In particular the thyroid gland and breast tissue are much more sensitive to radiation than other parts of the body and the rate of leukemia is much higher for radiation exposure compared to other known cancer causing agents. Another factor affecting cancer rates among exposed populations is age; children have a higher lifetime risk to a given exposure because there is more time for a cancer to develop after damage done by exposure.

5.6 Safety concerns

The energy available from 1 kg of $^{235}_{92}\mathrm{U}$ is roughly that found in 10^6 kg of coal or 14 000 barrels of petroleum. As a result, mining coal requires removing thousands of times more earth than mining for uranium. Based on historical data, this has resulted in approximately 40 times the number of accidental deaths for coal mining compared to uranium per MWh generated in the three decades before 2000 (Hirschberg 2004, Markandya and Wilkinson 2007). Burning coal also introduces heavy metals, sulfur and carbon dioxide into the atmosphere compared to nuclear power plants which are free of these emissions (except in the construction of the power plant itself). Burning coal concentrates the naturally occurring uranium and thorium isotopes in coal with the result that radiation exposure to coal ash near a coal plant may be as much 100 times that of the exposure from a neighboring nuclear power plant, although this exposure is very small for either type of plant compared to natural sources. Air pollution from burning fossil fuel (coal in particular) has a much larger impact because of the quantity of fuel being burned. It has been estimated that existing nuclear power plants have prevented some 1.84 million air pollution related deaths over their lifetimes as the result of reducing the amount of coal burnt for electricity (Kharecha and Hansen 2013). The general consensus of the number of deaths resulting from the Chernobyl nuclear disaster is 31 immediate deaths and some 4000 premature deaths due to cancer. Upper limits of the estimate of the number of future premature deaths due to radiation released from the Fukushima disaster are on the order of a few hundred. Based on these considerations, it would seem that statistically, nuclear power presents a far lower risk to health than traditional fossil fuel sources, even if nuclear disasters are included.

Although the real risks of nuclear energy compared to fossil fuel energy are statistically very low, nuclear reactor failures are spectacularly catastrophic. Two such failures, the Chernobyl disaster of 1986 and the Fukushima disaster of 2011 have been instructive for the future of nuclear power.

The Chernobyl event was a human caused disaster that occurred in a reactor in Ukraine. Operators at the plant were attempting to do an experiment to determine if the energy remaining in the plant as it was shutting down was sufficient to operate emergency equipment in the case of a power failure. Many of the safety and warning features of the plant were turned off during the experiment and a second, less well trained crew took over during the trial. The reactor design is also such that water evaporating from the core causes the chain reaction to increase in intensity, rather than decrease as it does for some other reactor designs where water is a moderator.

When the heat coming from the core began to increase very rapidly the operators could not move the control rods fast enough to slow the reaction and lost control of the plant in a matter of seconds. The heat from the core was high enough to separate hydrogen from water and ignite the hydrogen in a chemical explosion, blowing the top off the reactor, which was not enclosed in a containment building. The graphite core of the reactor then caught fire which carried radioactive material high into the atmosphere where it traveled thousands of kilometers. Approximately 350 000 people were eventually evacuated and an area of 2600 km^2 is now included in an exclusion zone around the plant due to high levels of radiation. This type of reactor is no longer being built.

The Fukushima disaster was the result of a once-in-a-thousand year tsunami which the reactor was not designed to withstand[12]. The plant began an automatic emergency shutdown as soon as the Earthquake which caused the tsunami was detected, long before the tsunami reached shore. Nuclear power plants, however, need on the order of a week to fully cool down, during which time water must continue to circulate through the reactor core or it will reach high temperatures and melt. Emergency diesel powered backup generators automatically came online to keep the cooling pumps running when power from the grid was cut off. Some of these generators, however, were located in a low area with a retaining wall that was not high enough to prevent flooding from the tsunami, estimated to be 13 m to 15 m high at the reactor site. When the tsunami flooded these generators the cooling pumps stopped and the core of one of the six reactors heated enough to cause hydrogen to form from the separation of water and explode, taking the roof off the building and sending radioactive material into the immediate surroundings, requiring the evacuation of thousands of people. Things were further complicated when sea water from helicopters was used to try to cool the reactors, flushing radioactive material into the surrounding land and ocean. As a result, the meltdown of the reactor core contaminated ground water at the location.

5.7 Nuclear fusion

Figure 5.1, showing the binding energy per nucleon, indicates that nuclei heavier than iron can decay into more stable nuclei, turning mass into energy thereby giving off energy. Modern nuclear reactors exploit this energy source to create steam which can be used to generate electricity. The figure also shows that for elements lighter than iron, fusing nuclei together should create more stable nuclei by lowering the binding energy per nucleon. This process, called *fusion*, also turns some mass into energy so that the combined nucleus weighs slightly less than the individual components and large amounts of energy are emitted per atom as is the case for decay processes. To overcome the electrical repulsion of the protons in two nuclei and fuse them, the atoms must be put under enormous pressure at high temperature. Two places where this occurs are in the Sun and in hydrogen bombs.

---

[12] Tsunamis causing damage or fatalities occur at a rate of about two per year; 10 m tsunamis occur at a rate of about 0.5 per 1000 years (NOAA 2016).

The actual process is complicated and takes many steps but can be summarized in the following reaction:

$$6\,{}^{1}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + 2\,{}^{1}_{1}\mathrm{H} + 2e^{+} + 2\nu + 2\gamma + 25\,\mathrm{MeV}. \tag{5.4}$$

Elements up to iron can be formed in a star if it is large enough so that the center of the star reaches appropriately high pressures and temperatures, and this depends on the size of the star. In about 4.5 to 5 billion years most of the hydrogen in the core of our Sun will be fused into helium and pressures and temperatures will not be high enough to cause further fusion. At this point the Sun will stop producing energy and undergo several changes in size and brightness due to the loss of fusion as a source of internal energy. Stars much larger than our Sun will fuse hydrogen into helium much faster and also be able to fuse helium into heavier elements (up to iron). Large stars collapse rapidly at the end of the fusion process which suddenly increases the pressure and temperature at the core causing an explosion (called a supernova), spewing these heavier elements into space. The elements, such as carbon, iron, etc, of which we are formed, all came from stellar fusion processes. Elements heavier than iron were formed during a supernova by chance collisions of lighter elements and for this reason are much more rare.

Although fusion research has been ongoing since the 1950s there are still substantial technical barriers to building a fusion power plant. The two basic methods being investigated are inertial confinement and magnetic confinement. In inertial confinement a small pellet of frozen hydrogen isotope is blasted by lasers from several different directions. The shock wave created by the sudden expansion of the surface of the pellet squeezes the core to pressures and temperatures high enough to attain fusion. The process is repeated rapidly and the energy given off is collected by a working fluid surrounding the containment vessel. In magnetic confinement magnetic fields squeeze a plasma of ions to high enough pressure and temperature to cause fusion. Both of these technologies are near the break-even point at which the energy produced by the reaction is approximately equal to the energy needed to cause and maintain the reaction. It remains to be seen if either of these processes can be scaled up sufficiently past the break-even point to become economically feasible and reliable as energy sources. If humans were able to achieve controllable hydrogen fusion a near perfect energy source would become available because, unlike fission, there are no dangerous fuels or waste products and the available energy is limitless for all practical purposes. Since fusion requires energy to maintain the process it is also inherently safe in that any failure in the reactor will cause the fusion process to stop.

Projects

1. Write a report on radioactive waste. Explain the US Nuclear Regulatory Commission definitions of types of waste and what is done with each type. Also discuss various plans to dispose of high level waste. What happens to it currently? What plans are underway to deal with radioactive waste on a long term basis?

2. The Sun emits energy at a rate of 3.8×10^{26} W. How much mass per second is being turned into energy? The Sun is projected to last another five billion years. How much of the Sun's current mass will be turned into energy by then?
3. Write a report on generation IV reactors. What are the different design types? Which ones are being built and where? What makes them inherently safe? Is there such a thing as inherently safe design (some critics think not)?
4. Write a report on the current state of fusion reactors. Include an analysis of the likelihood we will see commercial fusion reactors in the near future.
5. Use the values in table 5.4 to make a plot of dose versus excess cancer per person. The slope of this graph is a linear dose model prediction of the cancers per sievert. Use this figure to estimate the number of deaths caused by the average exposure of 3.6 mSv yr^{-1} in the US or some other country or region. Compare this prediction with other causes of death such as automobiles or heart attacks. Also compare this with the number of deaths at the average non-natural exposure rate of 0.6 mSv yr^{-1}.
6. Find information (a reputable web site is ok) to estimate your radiation exposure per year and comment on your findings.

References

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Halperin E C, Perez C A, Brady L W, Wazer D E and Freeman C 2007 *Perez and Brady's Principles and Practice of Radiation Oncology* (Philadelphia, PA: Lippincott, Williams and Wilkins)

Hafemeister D 2007 *Pysics of Societal Issues* (New York: Springer)

Hirschberg S, Burgherr P, Spiekerman G and Dones R 2004 Severe accidents in the energy sector: comparative perspective *J. Hazard. Mater.* **111** 57–65

IAEA (International Atomic Energy Agency) 2016 www.iaea.org/pris/ (Accessed: 1 September 2016)

Kharecha P A and Hansen J E 2013 Prevented mortality and greenhouse gas emissions from historical and projected nuclear power *Env. Sci. Technol.* **47** 4889–95

Lamarsh J R and Baratta A J 2001 *Introduction to Nuclear Engineering* (Englewood Cliffs, NJ: Prentice Hall)

Markandya A and Wilkinson P 2007 Electricity generation and health *Lancet* **370** 979

NRC (National Research Council) 2006 *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2* (Washington DC: National Academies Press)

NOAA (National Ocean and Atmospheric Administration) 2016 NOAA Global Historical Tsunami Database https://www.ngdc.noaa.gov/hazard/tsu_db.shtml (Accessed 2 April 2017)

World Nuclear Association 2016 www.world-nuclear.org/information-library/nuclear-fuel-cycle/uranium-resources/supply-of-uranium.aspx (Accessed: 1 September 2016)

Chapter 6

Renewable energy

All renewables together currently account for less than 14% of the global primary energy supply. The majority of renewable energy used today is in the form of biomass, burned for heating and cooling in rural regions of less developed countries[1]. Most of the rest is in the form of bio-fuels or hydroelectric dams. Although they are the fastest growing renewable resources, wind and solar still make up only miniscule portions of the global energy supply. This chapter surveys the current status of renewable energy and speculates on the prospects of a renewable energy future.

6.1 Hydropower

Globally most dams are used for irrigation but some 16% of all dams in the world over 15 m in height are used for electric generation (ICLD 2016). According to the International Hydropower Association (IHA), current world capacity for generating electricity from these hydropower plants is 1212 GW (IHA 2016). In 2015, 37.4 EJ of electricity from hydroelectric sources were consumed globally which constitutes 7% of the world's total supply of energy and 16% of its electricity (BP 2016). A typical hydroelectric plant can be seen in figure 6.1. Hydroelectric dams are also used for energy storage, which will be discussed in chapter 7. Hydroelectric plants are attractive as an energy source because they are the least expensive renewable energy and the least polluting. The lakes of water stored behind dams can also be used for other purposes, for example irrigation, drinking water and recreation.

The five top global consumers of hydroelectric power (in order) are China, Brazil, the US, Russia and Canada. Norway generates nearly all of its electricity from hydropower, Nepal and Brazil more than 90%, New Zealand 55%, Canada 60% and Sweden generates about half of its electricity from hydroelectric (IHA 2016). As can

---

[1] Although this may be changing. The UK currently supplies 10% of their electricity from wood chips made from compressing the remaining branches after lumber is harvested in the Southeastern US.

doi:10.1088/978-1-6817-4493-3ch6

Figure 6.1. Hoover dam is located on the border between Nevada and Arizona in the US southwest. In 2015 the dam generated 3.6 TWh = 1.3×10^{-2} EJ. (Photographs taken by the author.)

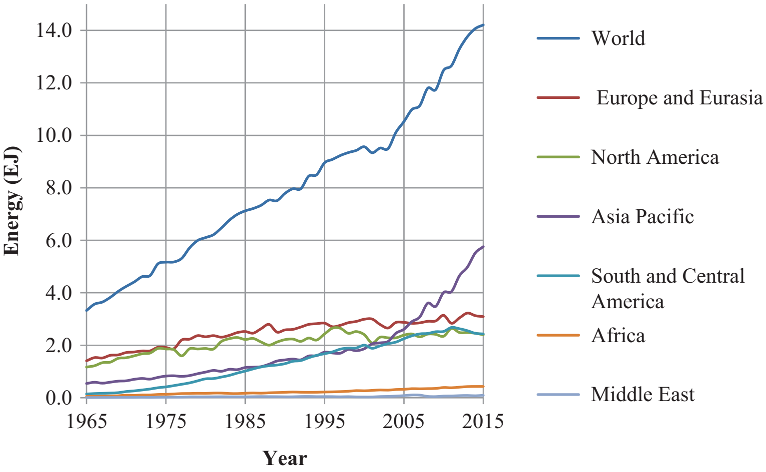


Figure 6.2. Global hydroelectric energy consumption (BP 2015).

be seen in figure 6.2, Asia (China in particular) has seen a sharp increase in the use of hydroelectric since 2000. In Europe and the US most major water flows that are economically feasible for hydroelectric dams have been exploited[2] and in fact the recent trend in the US is to tear down smaller dams and restore these rivers to their natural states as closely as possible. Hydroelectric has decreased from 40% of the US electrical supply in the 1940s to 8% today, largely because it has not expanded while other sources have been brought online to meet demand. There is still a large untapped potential for hydroelectric in Asia, Africa and South America (Kumar *et al* 2011, IPCC 2012).

Hydropower has long been considered the most environmentally benign of the renewable energy resources but no power source is completely free of environmental

---

[2] North America has tapped 39% of available hydroelectric resources, Europe 53%, Asia 20%, Africa 8% and Latin America 26% (IPCC 2012).

impact. The process of making cement is a significant source of carbon dioxide and decaying vegetation in the lakes behind a dam can also emit substantial amounts of carbon dioxide and methane. The disruption of natural habitats has been recognized as a significant outcome of dam building and there is the potential for millions of people being displaced in locations where population density is high, such as occurred during the construction of the Three Gorges Dam in China. Large dams also create the risk of geological shifts and earthquakes due to the weight of water behind the dam. In dryer climates, evaporation from the lake behind a dam may result in significant loss of water intended for downstream usage after release from the dam. The economic gains from building hydroelectric dams decreases with the size of the dam, with the result that it is generally less cost effective to dam small tributaries. These environmental and economic issues have already slowed dam development in North America and Europe and are likely to eventually limit the amount of energy derived from dams globally.

The gravitational potential energy of a mass, m at height, h is $U = mgh$, where g is the acceleration of gravity. Consider a rectangular block of water passing over a dam which has a volume Ax where A is the frontal surface area of the block (e.g. the cross sectional area of the pipe of water going to a generator or the cross sectional area of a spillway) and x is some arbitrary length. The mass of this block of water is $m = \rho V = \rho Ax$ where ρ is density and the potential energy it has is $U = \rho Axgh$. Dividing by time gives the power available and using $v = x/t$ for the velocity of the water we can write

$$P = \frac{U}{t} = \rho Avgh. \tag{6.1}$$

This expression is called the hydraulic head and is the main source of energy for most large scale commercial dams. This can also be written in terms of the flow rate, $Q = Av$, as

$$P = \rho Qgh, \tag{6.2}$$

where Q is measured in cubic meters per second.

Water in motion also has kinetic energy, $\mathrm{KE} = \frac{1}{2}mv^2 = \frac{1}{2}\rho Axv^2$ and the kinetic power of water in motion is therefore $P = \frac{1}{2}\rho Av^3$. Undershot waterwheels and underwater turbines tap this power source, rather than the hydraulic head. The water must flow away from the bottom of the waterwheel or dam and so it still has some residual kinetic energy which cannot be tapped. If the velocity changes from v_i to v_f then the power generated is

$$P = \frac{1}{2}\rho A\left(v_\mathrm{i}^3 - v_\mathrm{f}^3\right). \tag{6.3}$$

The total power available from water in a dam of height h is therefore

$$P = \rho Qgh + \frac{1}{2}\rho A\left(v_\mathrm{i}^3 - v_\mathrm{f}^3\right), \tag{6.4}$$

where v_i is the velocity of the water as it reaches the dam and v_f is the velocity as it flows away from the bottom of the dam. The same equation can be used for moving water in a river, the ocean or due to tides by dropping the first term.

6.2 Wind

Estimates of the amount of available wind energy globally over land are as high as 30 000 EJ, depending on the height, size and design of the windmill (WEC 2013). Even if only a small fraction of this estimate is captured it could make a significant contribution to global energy consumption and several countries, Denmark for example, already receives a substatial proportion of its electricity from wind. Global wind energy consumption is shown in figure 6.3.

Most modern wind turbines are either vertical axis (propeller type) or horizontal axis (eggbeater or Darrieus type), although there are scores of other designs. The modern horizontal type usually has two or three blades, although turbines with a single blade and a counterweight have been built. Large modern windmills can have blades up to 63 m in length and hub heights up to 135 m. A horizontal axis blade from a small windmill is shown in figure 6.4. For all wings, including turbine blades, the wind passes over a surface curved on one side creating a lift on the blade. A drag or friction force is also present and the blade's motion is the result of these two forces. A detailed explanation of how the lift and drag forces act on a wing or wind turbine blade can be found in Denker (2014).

Like many other renewable energy resources wind provides an intermittent supply of energy which varies during the day and seasonally by as much as 30%. Generally wind speeds lower than three to four meters per second are not useful for a large modern turbine and a constant maximum power output occurs for winds above 10 m s^{-1}. Too much wind can also be a problem as the torque on the tower as a result of wind drag increases significantly in storms or high winds. Most turbines

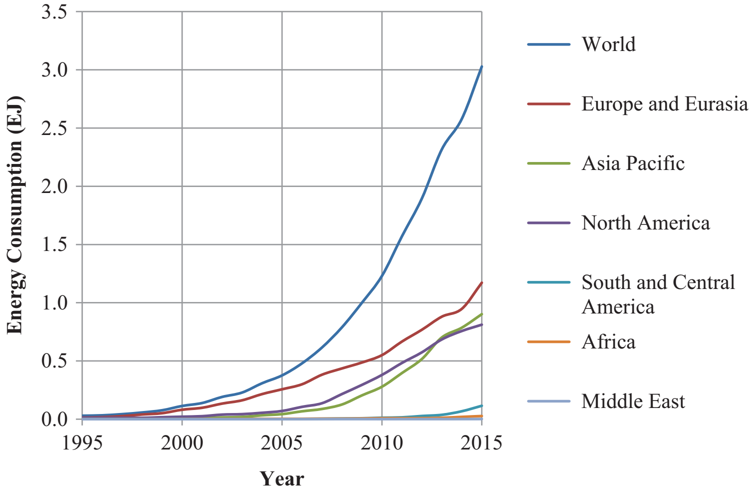


Figure 6.3. Global wind energy consumption (BP 2016).

Figure 6.4. A single wind turbine blade, 20 m long, from a small wind turbine. (Photograph taken by the author.)

cease generating electricity at wind speeds above 25 m s^{-1} and the blades are allowed to spin freely to avoid damage to the windmill. Windrose diagrams which show the direction and strength of the wind at specific locations in the US can be found at (NRCS 2016). Wind maps for the US are available at (NRELa 2016). Global wind maps are available at (SWERA 2016).

The power of a windmill scales as the cube of velocity (as shown below) which makes it very sensitive to changes in wind speed. If the average wind in a particular location is 7 m s^{-1} but the wind fluctuates so that it has velocities of 3 m s^{-1}, 7 m s^{-1} and 11 m s^{-1} for equal amounts of time, the average power will be proportional to $((3\ \mathrm{m\ s^{-1}})^3 + (7\ \mathrm{m\ s^{-1}})^3 + (11\ \mathrm{m\ s^{-1}})^3)/3 = 1.6(7\ \mathrm{m\ s^{-1}})^3$ or $1.6v_{\mathrm{ave}}^3$. Having a wind speed 4 m s^{-1} higher for one third of the time and 4 m s^{-1} lower for one third of the time increases the power output by 1.6 more than if the speed stayed constant at the average 7 m s^{-1}.

It has been found empirically that wind speed scales with height to the 1/7 power near the Earth's surface which means that if the wind at 10 m has a speed of $v_{10\mathrm{m}}$ then the speed of the wind at a height of 60 m will be $v_{60\mathrm{m}} = (60/10)^{1/7}v_{10\mathrm{m}} = 1.3v_{10\mathrm{m}}$. A windmill taking advantage of a 7 m s^{-1} wind at a height of 10 m will have $(1.3)^3 = 2.2$ times more power output if raised to 60 m. The hub of a horizontal axis windmill should therefore be positioned as high as practically possible in order to take advantage of the cubic velocity scaling. Because of the scaling with height, blades must be able to withstand a significant difference in force from one end to the other in order to withstand the difference in forces at different heights as the blades rotate.

Wind resources are widely distributed with much stronger winds occurring far out to sea compared to wind speeds over land. Even on land wind resources are sometimes distant from population centers, for example the Patagonia region of

South America has some of the highest winds of South America, but also the lowest population density. Because of transmission loss in electrical lines, sources remote from population centers are less likely to be economically competitive. Some five countries have nearly half of the global wind potential but have only 10% of the global population (Moriarty and Honnery 2012). Individual windmills do not take up much land but must be sparsely distributed so as not to interfere with each other. As a rough guide, a spacing of five turbine diameters reduces the power in a turbine behind it by 30%, a 10% reduction is seen if they are spaced at distances 10 times the diameter of the windmill (Hafemeister 2007).

There is some concern that windmills will contribute to the rate of birds killed by modern society and this is certainly true, particularly if the windmill is located on a bird migration route. However, estimations of birds killed by various sources indicate that cats, either outdoor pets or feral, currently kill as much as 2000 times more birds than windmills and birds flying into the windows of tall buildings is twice that. Power lines, cars, pesticides, high tension wires and communication towers each contribute far more avian fatalities than windmills. An increase in the number of windmills will increase the number of bird killed, but estimates of the bird kill per GWh from nuclear and fossil fuels plants are twice and 30 times higher, respectively, than for windmills (Sovacool 2012).

The kinetic energy in wind can be written as $\mathrm{KE} = \frac{1}{2}mv^2$ where m is the mass of a hypothetical block of wind traveling at speed v. The power contained in this quantity of wind with vertical surface area A and length x moving at velocity v is then

$$P = \frac{1}{2}\rho A v^3, \tag{6.5}$$

where ρ is the density of air in kg m^{-3}. The air does not stop at the blade and will exit with some lower speed so that not all the energy can be captured (as was the case for water flow).

For wind it is more convenient to work with mass flow in kg s^{-1} rather than the flow rate, Q, in m^3 s^{-1} as was done for hydropower because air can compress. Although the volume and pressure of a parcel of air may change as it passes a turbine, the total mass of air remains fixed. The mass flow can be defined as

$$\delta m = \rho A v_{\mathrm{ave}}, \tag{6.6}$$

where $v_{\mathrm{ave}} = \frac{1}{2}(v_{\mathrm{i}} + v_{\mathrm{f}})$, and v_{i} and v_{f} are the wind speeds before and after passing the turbine. The change in kinetic energy per second of this parcel as it passes the turbine is $\frac{\Delta \mathrm{KE}}{\mathrm{time}} = \frac{1}{2}\delta m(v_{\mathrm{i}}^2 - v_{\mathrm{f}}^2)$, which is the energy captured by the turbine per second. Putting equation (6.6) into this expression gives the power generated by the turbine as

$$P = \frac{\Delta \mathrm{KE}}{\mathrm{time}} = \frac{1}{2}\rho A v_{\mathrm{ave}}\left(v_{\mathrm{i}}^2 - v_{\mathrm{f}}^2\right). \tag{6.7}$$

After some rearranging and substitutions we have

$$P = \frac{1}{4}\rho A v_{\mathrm{i}}^{3}\left(1 + \frac{v_{\mathrm{f}}}{v_{\mathrm{i}}} - \left(\frac{v_{\mathrm{f}}}{v_{\mathrm{i}}}\right)^{2} - \left(\frac{v_{\mathrm{f}}}{v_{\mathrm{i}}}\right)^{3}\right) \tag{6.8}$$

for the power generated.

Setting the derivative respect to $\frac{v_{\mathrm{f}}}{v_{\mathrm{i}}}$ of the power equation equal to zero will give the maximum power possible and this yields $\frac{v_{\mathrm{f}}}{v_{\mathrm{i}}} = \frac{1}{3}$. Putting this into equation (6.8) gives

$$P_{\max} = \frac{16}{27}\left(\frac{1}{2}\rho A v_{\mathrm{i}}^{3}\right). \tag{6.9}$$

The total power available is $\frac{1}{2}\rho A v_{\mathrm{i}}^{3}$ (equation (6.5)) so the ratio $\frac{16}{27} = 59\%$ gives the upper limit to the efficiency of a wind turbine which occurs when the wind speed drops by two thirds after passing the turbine. This limit is known as the Betz limit after its discoverer, Albert Betz.

Modern horizontal axis two and three blade windmills have achieved efficiencies of 45% to 50% at wind speeds of 6 m s^{-1}, but normal windmill efficiencies and efficiencies of other designs are often half of this. Over time windmills accumulate dirt and insects on the blades which can reduce the power collected by 50%. Periodic cleaning of windmills is difficult due to the large size and heights of large modern windmills. Windmills must also be designed to take into account the not insignificant chance of lightning strikes due to their height.

6.3 Biomass

The trend over the last decade has been to consider biomass, bio-fuels and municipal landfill waste as separate resources, but until recently all of these were included in the term *biomass* which applied to any organic material that could be used as a fuel or energy source. Under the more generic term are included wood and wood products such as charcoal and black liquor (processed wood pulp); wood residue (sawdust); crop residue (straw, stalks); animal manure; sewage sludge; municipal waste; yard waste; landfill gas; biogas; and bio-fuels such as ethanol and biodiesel. Crops grown for the production of bio-fuel include switch grass, corn and soybean, in the US, *Miscanthus* in Europe, cassava in parts of Southeast Asia and Africa, and sugarcane in Brazil. Fuels from algae are under investigation but are not yet a significant fuel source.

Currently biomass supplies about 50 EJ of energy to the world which is about 10% of the global energy demand[3]. Most of this comes from wood, wood products, agricultural and municipal waste, and is used for cooking and heating in poor rural parts of the world (WEC 2013). Global bio-fuel production was more than 3 EJ in 2015, as shown in figure 6.5. In the US in 2011 wood and bio-fuels made up 43% of

---

[3] Global energy use is currently at around 550 EJ.

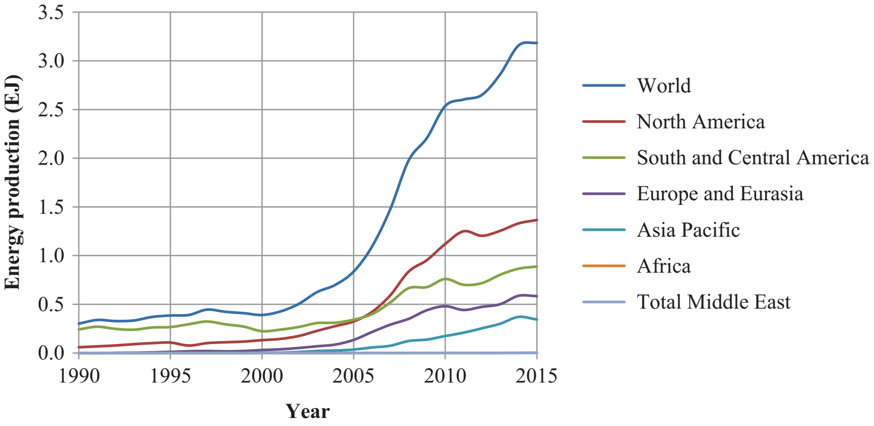


Figure 6.5. Global bio-fuel production (BP 2016).

the renewable energy used and 4% of all primary energy[4]. Brazil is the second leading producer of bio-fuel (ethanol from sugarcane) in the world and has a mandated ratio of 25% ethanol content of motor vehicular fuels at the pump.

The chemical reactions that produce bio-fuel from a raw biomass feedstock depends on the biomass. The sugars in sugarcane and sweet sorghum can be converted into ethanol and butanol by anaerobic fermentation. To be used as a fuel the starch in maize (corn), cassava or other fuel crops must first be converted to sugar by a saccharification process which can be chemical or biological. Switch grass, poplar, *Miscanthus*, corn stover, bagasse (sugarcane residue), forest trimmings, yard waste and construction debris are composed mostly of cellulose which can be either gasified at high temperature, or treated by pyrolysis or the Fischer–Tropsch process to form complex hydrocarbons from which fuel can be obtained. There has been some progress in using cellulosic fermentation which employs special strains of bacteria to convert waste biomass directly into complex hydrocarbons, but this is not commercially viable yet. Palm oil, soy oil and commercial waste oil can be treated by a chemical process called trans-esterification to form biodiesel fuel. As an alternative any of these primary feedstocks can be burned to produce steam for electricity generation with varying degrees of thermal efficiency.

Some estimates put the global potential for biomass as high as 1500 EJ per year, but this seems unlikely given the low overall efficiency of plant photosynthesis. Although the photosynthesis process itself is very efficient, plants are at best only 8% efficient in converting solar energy into a form which can be harvested (Zhu *et al* 2008)[5]. Managed energy fuel crops such as tree plantations or sugarcane grown with modern farming techniques are estimated to generate between 0.4 and 1.2 W m^{-2},

---

[4] Approximately one third of the corn grown in the US is used to make ethanol which is added to gasoline, and one third of the wood harvested in the Southeastern US in 2013 was used to make wood pellets for fueling electric power plants, mostly in the United Kingdom.

[5] As mentioned in chapter 5, typical plant efficiencies for turning sunlight into a usable fuel source are around 1%.

with crop residues and wood from unmanaged forests at about half this much (Smil 2005, Weisz 2004). For comparison, commercially available PV solar cells are currently 15% to 25% efficient in converting sunlight into electrical energy, a factor of 10 to 20 times better than fuel crops per square meter of light absorbed.

As a sample calculation about the limits of bio-crops as a fuel source consider the total arable land in the US which is about 1.6×10^6 km^2. Using 1.2 W m^{-2} as a power output for a fuel crop we have a power output of $1.6 \times 10^6\ \mathrm{km}^2 \times 1.2\ \mathrm{W\ m}^{-2} = 1.9 \times 10^{12}$ W. If this could be maintained the entire year, including winter, it would amount to around 61 EJ if all of the energy could be harvested with no energy invested. Using an EROEI (see chapter 4) of 1.6 for ethanol from corn we would need about 38 EJ input energy to create this fuel, so the net contribution to the energy supply is only 23 EJ. If sugarcane, with an EROEI of 10, could be grown everywhere in the US[6] only 6.1 EJ would be needed as input and 55 EJ would be contributed to the net energy supply, a little more than half the total primary energy consumed in 2016 in the US. Algae, particularly if it can be genetically modified to produce more plant oils, may end up with a higher EROEI than sugarcane and has the advantage that it can live in salt or brackish water, does not take up arable land used for crops or livestock and is not part of the human food supply.

Over the past century the US has been able to increase crop yield by a factor of five using increased fertilization, hybrid plants, improved pesticides and modern farming methods. It is quite possible this trend will continue and the spread of these techniques to farming in less developed areas of the world will undoubtedly improve crop yields. Some recent experiments with genetic engineering of the respiratory system of plant leaves have been able to increase plant yield by 15% in fluctuating light conditions (Kromdijk *et al* 2016). But land needed for food crops, the need to preserve natural habitats and the need for land for human habitation would preclude the use of the entire biomass of the Earth for fuel. Given these limitations and the low efficiency of plant growth, it seems unlikely that bio-fuel will supply all of the world's energy demand, although it could play a significant role, especially as a portable energy source (see Smil (2015) for a comprehensive analysis of global biomass supplies).

Two positive aspects of bio-fuels are that they are a renewable resource and plants undergoing photosynthesis absorb carbon dioxide to make sugar and oxygen as delineated in this generic chemical reaction

$$6\mathrm{CO_2} + 6\mathrm{H_2O} \rightarrow \mathrm{C_6H_{12}O_6} + 6\mathrm{O_2}. \tag{6.10}$$

One can argue that the net carbon dioxide flux to the environment in the combined process of photosynthesis and burning the sugar as an energy source should be approximately zero. It is not entirely clear if this is actually the case since there is some CO_2 added to the atmosphere during the planting and processing phases due to the low EROEI for bio-fuels. Whether bio-fuels are truly carbon neutral remains to be seen (Korner 2017).

---

[6] It is the relatively high energy balance for the production of ethanol from sugarcane that makes the use of ethanol in Brazil successful; the low energy return of ethanol from corn makes it much less attractive as a bio-fuel.

In contrast to bio-fuels, municipal waste as an energy resource undeniably has a positive energy return. Plastic, some 10% of the modern waste stream, has an energy content of 42 MJ kg^{-1}, close to the 47 MJ kg^{-1} contained in gasoline. The energy content of paper, another significant waste stream, is 40 MJ kg^{-1} and wood scrap 15 MJ kg^{-1}. In high and middle income countries paper, organics and plastic, all high energy content material, make up more than 75% of the municipal waste. In less developed regions these three categories make up 50% of the municipal waste (WEC 2013). Global municipal waste is on the order of 1.3 billion tonnes per year, which amounts to 0.5 kg per capita per day, and this is expected to increase as the world's population becomes more urban. For municipal waste with an average energy content of 10 MJ kg^{-1} the global primary energy available would be 13 EJ.

Generation plants using municipal waste as fuel have the advantage that the fuel resource has in effect a negative cost since waste removal generally costs money. A negative aspect of burning municipal waste is that heavy metals and other chemicals which are not consumed in the burning process can be emitted into the atmosphere. Currently the majority of electric generating plants using municipal waste as a fuel co-produce heat for manufacturing and domestic use (WEC 2013). About 15% of municipal waste in the US is burned using energy recovery systems for either heat or electricity according to US Department of Energy figures. It should be mentioned that a side effect of efficient recycling programs is that they would likely reduce the energy content of municipal waste streams.

6.4 Geothermal

The Earth's interior is at a temperature near 4500 °C as a result of radioactive elements and gravitational pressure at the core. A very thin layer of relatively cool rock at the surface of the Earth floats on a layer of mantle with temperatures between 500 °C and 1000 °C compared to average surface temperatures of less than 40 °C. This temperature gradient results in a heat flux from the interior of 0.06 W m^{-2} on average, although there are many places where the mantle is closer to the surface and the flux is much higher. The total energy flux per year amounts to 950 EJ. There are three basic approaches to utilizing this resource; deep mining, thermal hotspots and heat pumps.

The large temperature difference between the surface and the mantle would make for a very efficient heat engine providing an enormous energy source and attempts to drill down to the mantle have been made. Under the ocean the Earth's crust is 5 km to 10 km thick and under continents it can be 30 km thick. Wells of 10 km depth are technically feasible, although costly, but there has yet to be a successful attempt to reach the mantle.

Geological hot spots occur at the boundaries of tectonic plates which are slowly moving pieces of the Earth's crust floating on the mantle below. Volcanoes and earthquakes occur in the geologically active regions where these plates grind together or slide under each other. At these boundaries hot molten rock at temperatures of 1000 °C may be within two or three kilometers of the surface. If water from the surface reaches these heat sources it becomes steam under high

Figure 6.6. Geothermal hot springs in Costa Rica. Air temperature about 8 °C. (Photograph taken by the author.)

pressure and cause hot springs (figure 6.6) and geysers on the Earth's surface. These naturally occurring steam generators have been used to make electricity since the early 1900s.

Most modern geothermal plants pump water or other working fluid down into a thermal hotspot to generate steam to drive electrical turbines or to use for heating and industrial processes. Thermal efficiencies for temperatures below 150 °C are not considered sufficient for generating electricity and most modern generating plants use wells that go down one to two kilometers where temperatures are 250 °C to 350 °C. The Brayton or Rankine cycles, discussed in chapter 3, apply to a steam driven turbine operating between these temperatures and the temperatures at the surface. Global geothermal capacity is currently about 0.37 EJ of electricity and 0.88 EJ of heat per year. These resources are strongly site dependent, however. For example Iceland is able to supply 25% of its electricity and 85% of its heating needs from geothermal because of the geological features underlying the country, but few other locations have this quantity of geothermal resource. Combined consumption of geothermal, biomass (not including bio-fuels) and other renewables are shown in figure 6.7.

A third way to use the heat flux from the Earth which will work at non-geologically active locations is the heat pump, as discussed in chapter 3. In regions where there is not a geological hot spot the steady heat flow from the Earth's core maintains a constant temperature at a given depth. The exact temperature at a particular depth depends somewhat on local geological features, but can be used as either a heat sink to store heat from a building in summer or, more commonly, a heat source from which to extract heat during the winter. A mechanical pump circulates a heat transfer fluid through pipes buried in the ground and this liquid

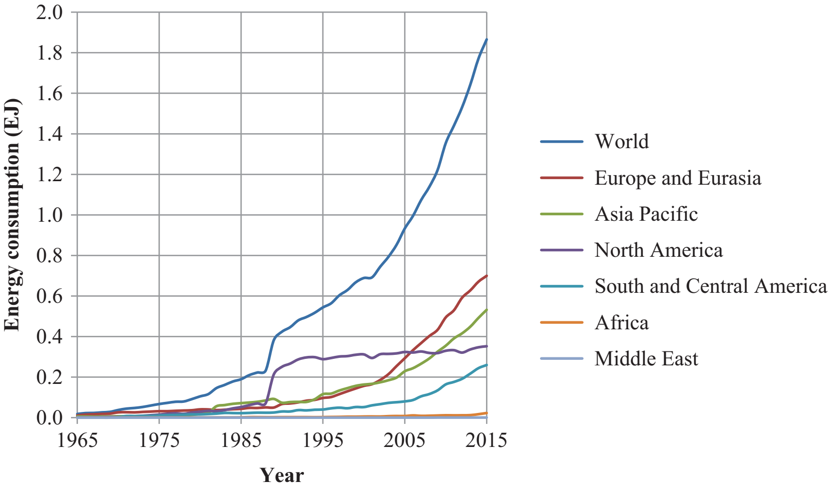


Figure 6.7. Global geothermal, biomass (not including bio-fuels) and other renewable (not including wind or solar) energy consumption (BP 2016). Geothermal accounts for about half the totals shown.

transports heat in the desired direction. Because the pump is simply moving heat, less primary energy is needed compared to burning gas in a home gas furnace for heating.

6.5 Solar

Passive solar energy storage, as seen in figure 6.8, has been used for millennia. In fact there are many surface locations around the world which stay close to the same temperature year round due to the high thermal capacity of stone or water surrounding the location. It is also the case that domestic hot water solar systems (mostly in warmer climates) supply nearly ten times the solar energy used globally than current large scale solar systems. In this section we examine PV and modern large scale thermal solar energy sources; storage of various kinds of solar energy are examined in the next chapter.

The amount of energy reaching the top of the Earth's atmosphere is known as the solar constant, S_0, and is measured to be 1366 W m^{-2}. The target presented by the Earth to rays coming from the Sun is circular so the total energy per second striking the Earth is $\pi R^2 S_0$. This power is spread out over the area of half of a sphere so the energy reaching the top of the Earth's atmosphere per second averaged over all latitudes is

$$\frac{\pi R^2 S_0}{\frac{1}{2} 4\pi R^2} = \frac{1}{2} S_0 = 683 \text{ W m}^{-2}. \tag{6.11}$$

The Earth's albedo is an average of 0.30, indicating that 30% of the incident energy is reflected back into space on average. The amount of solar radiation reaching the Earth's surface averaged over all latitudes is then

Figure 6.8. Site chosen for natural passive solar heating, Mesa Verde, Colorado, US. (Photograph taken by the author.)

$$\frac{1}{2}(1-a)S_0 = 478 \text{ W m}^{-2}, \tag{6.12}$$

not taking the angle of incidence or time of day into account.

The flux reaching the surface is also very dependent on the local climate, varying by as much as 70% for areas which have significant average amounts of cloud cover compared to arid, high altitude locations. The Equator gets more power and the poles less because of the angle of the Sun's rays relative to the Earth's surface. The atmosphere is generally transparent to visible frequencies but some other frequencies are blocked by water vapor and other gases in the atmosphere. Intensity is further reduced by as much as 15% from the fact that at low angles (high latitudes or sunset and sunrise) the sunlight passes through a larger amount of atmosphere to reach the surface.

For the purpose of knowing how much energy is arriving at a particular location on the earth's surface and time, known as the *insolation*, we can start with the unreflected part of the solar constant which would be about 1000 W m^{-2} for a square surface directly facing the Sun when it is directly overhead at the Equator with no clouds[7]. The insolation, I, for a particular time and location of the Earth can then be calculated by

$$I = S\cos(Z), \tag{6.13}$$

---

[7] 1000 W m^{-2} is equivalent to 24 kWh m^{-2} per day, a common industry unit termed one Sun.

where

$$Z = \cos^{-1}(\sin\phi \sin\delta + \cos\phi \cos\delta \cos H) \tag{6.14}$$

and $S = 1000\ \mathrm{W\ m^{-2}}$. Here ϕ is the latitude of the location, δ is the declination which varies in the Northern Hemisphere from 23.5° on the summer solstice (21 or 22 June) to −23.5° on the winter solstice (21 or 22 December) and H is the hour angle. The hour angle is given by $H = 15° \times (\mathrm{time}-12)$, where *time* is the hour of the day counting from midnight.

As an example, at a latitude of 60° at noon on the equinox (21 or 22 September) the hour angle is $H = 15° \times (\mathrm{time}-12) = 0$. On the equinox the declination is exactly halfway between 23.5° and −23.5° and so is also zero. We have $Z = \cos^{-1}(\sin 60 \sin 0 + \cos 60 \cos 0 \cos 0) = 60°$. The insolation for a latitude of 60° at noon on the equinox is therefore $I = S\cos(Z) = 1000\ \mathrm{W\ m^{-2}}\cos(60) = 500\ \mathrm{W\ m^{-2}}$.

From these and similar calculations the daily average power reaching the Equator is between 500 W $\mathrm{m^{-2}}$ and 550 W $\mathrm{m^{-2}}$ (Sorensen 2000). This is also the approximate insolation for much of the Northern Hemisphere during the summer months (or Southern Hemisphere in winter) assuming there are no clouds blocking the Sun. In the winter months the insolation average in the Northern Hemisphere drops to between 150 W $\mathrm{m^{-2}}$ and 200 W $\mathrm{m^{-2}}$ at a latitude of 40° due to the change in angle between the Sun and the surface of the Earth. For middle and low latitudes, which constitutes the areas where most of the world's population lives, the average daily power reaching the Earth's surface can be estimated to be about 250 W $\mathrm{m^{-2}}$ if cloud cover is minimal (insolation maps for the US are available from NREL (2017)).

Solar collector systems are often rated in terms of their available *peak power* which usually assumes an incoming solar flux of 1000 W $\mathrm{m^{-2}}$ at noon with the panel directly facing the Sun. The actual power depends not only on weather conditions but also on the angle of the Sun and the latitude where it is observed, as shown in equation (6.13).

The average square footage of a house in the US is around 2000 $\mathrm{ft^2}$ which is very roughly 200 $\mathrm{m^2}$. Using the 250 W $\mathrm{m^{-2}}$ average insolation and a rooftop solar panel of 200 $\mathrm{m^2}$ the available power would be $250\ \mathrm{W\ m^{-2}} \times 200\ \mathrm{m^2} = 50\ \mathrm{kW}$, assuming all the energy is captured. A 20% efficient PV solar cell array (discussed below) would collect 10 kW. Average per capita electricity use in the US is 1.4 kW so this would be enough to supply a family of seven or an apartment building with the same rooftop and seven individual dwellings of 200 $\mathrm{m^2}$.

Based on an average power of 250 W $\mathrm{m^{-2}}$ the total solar energy reaching the world's land mass is 8.5×10^5 EJ annually with another 1.7×10^6 EJ reaching the surface of the ocean. Using 20% efficient solar cells and transmission, storage and conversion losses of 50%, an area of solar cells around 602 000 $\mathrm{km^2}$ would be needed to supply the entire world's energy needs. Using the same assumptions, an area about the size of North Carolina covered with solar cells would supply the entire electrical supply of the US. Estimates of the percentage of a country's land area needed to supply *all* the country's energy is shown in table 6.1 for a few select countries.

Table 6.1. Annual energy use of select countries and the percent of land mass of the country covered with solar cells or windmills which would supply all of the annual energy use. The solar figures assume 150 W m^{-2} arriving at the surface during winter for 9 h/day and 500 W m^{-2} arriving during summer for 15 h/day (40° latitude average figures), half the time (clouds bock half the incoming radiation). The solar cells are 20% efficient and transmission is 50% efficient. Wind is based on the 0.5 km^2 $(MW)^{-1}$ production average of existing windmill farms and a 50% transmission loss.

Country	Annual energy consumption (EJ)	Land surface needed for all wind (%)	Land surface needed for all solar (%)
Argentina	3.7	1.0	0.5
China	128.1	9.6	4.7
Denmark	0.7	11.7	5.7
Egypt	3.7	2.6	1.3
France	10.2	11.4	5.5
Ghana	0.2	0.6	0.3
Japan	19.1	36.3	17.5
Russia	28.3	1.2	0.6
UK	3.9	11.6	5.5
US	97	7.1	3.4

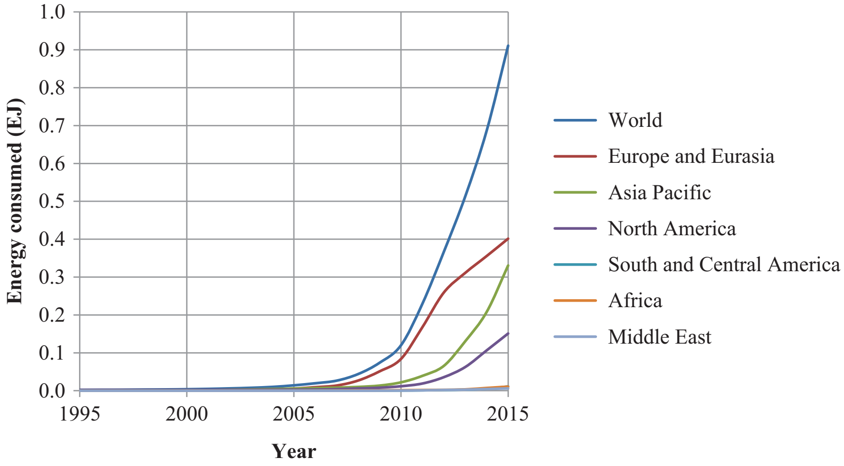


Figure 6.9. Global solar energy consumption (BP 2016).

It should be clear from the above calculations that there is more than enough solar energy to supply all of the global demand for energy by a factor of at least 1500. Global solar energy consumption, however, is still quite low, as can be seen in figure 6.9. This is likely to change significantly in the next decade as the price of solar energy has now dropped to be level with other energy sources such as natural gas and wind without tax incentives (see the discussion of table 6.3 below).

The two large scale types of solar energy collection, solar thermal and PV, are discussed in the next section. Currently PV technology is much more advanced commercially than solar thermal and is the largest growing renewable energy source.

6.5.1 Photovoltaic

In chapter 2 the mechanism of LED lighting was explained. Two crystals with different energy bands and band gaps are placed in contact so that electrons from a higher energy band in one crystal will drop across an energy band gap to a lower energy band in the other crystal, giving off energy. This electromagnetic energy, in the form of a photon, is $E = hf$, where h is Planck's constant and f is the frequency of the photon. The process requires input current so that there is a steady supply of electrons changing energy levels and giving off photons. The reverse process, using basically the same crystals and physical arrangements can also occur, and this arrangement is called a photovoltaic (PV) cell. Incoming photons with enough energy can create an electron–hole pair at the junction of p- and n-type semiconductors. Free electrons are formed on the n-side by boosting electrons over the band gap into the conduction zone where they can move as a current. On the p-side boosting electrons out of the valence band leaves holes which act as positively charged particles that are also now free to move. Electrons created on the n-side of the junction repel each other and flow away from the junction while positive holes flow away from the junction on the p-side. The flow of electrons constitutes an electric current and the energy of the moving electrons can be used in an external device, represented by the resistor, R, in figure 6.10. In the figure the conventional current, I, in amperes flows to the right since conventional current is labeled to flow opposite the direction of the electron motion.

The band gap limits the efficiency of a solar cell since only photons with frequencies high enough to have an energy large enough to boost an electron over the band gap are useful. Only photons with $E > E_{\text{band}}$ where E_{band} is the energy of the band gap can be captured. If the photon has more than enough energy to cross the gap this extra energy does not produce more current but instead turns up as excess thermal energy in the photocell. Some energy is also lost by unwanted electron–hole recombination before they have a chance to move away from the junction. This loss can be calculated from the open circuit voltage; the voltage across

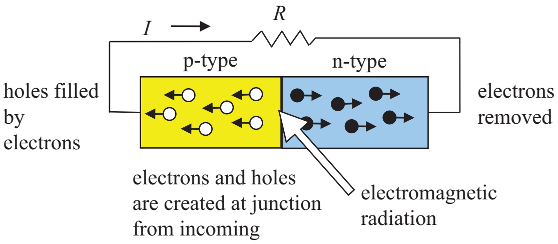


Figure 6.10. Schematic of a generic solar cell. Photons (electromagnetic radiation) cause electrons to jump across the energy band gap at the junction of a p–n diode. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

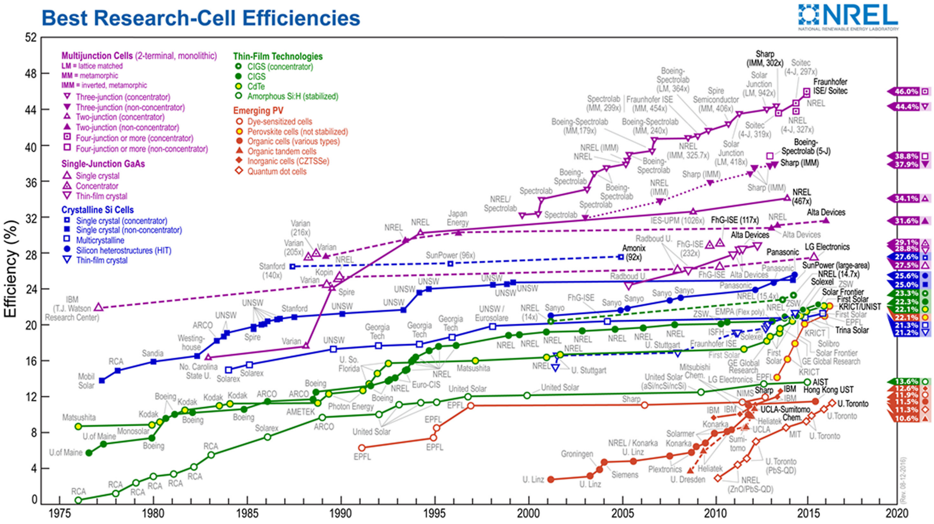


Figure 6.11. Laboratory efficiencies of various prototype PV cells (NRELb 2016).

the PV cell when it is not connected to a circuit. The ratio of the open circuit voltage to the band gap voltage is the percentage of energy lost to recombination which also turns up as heat. The energy in sunlight peaks in the visible part of the spectrum so most solar cells are designed to operate as close to this part of the spectrum as possible but they do so at a narrow range of frequencies.

In the case of PV cells efficiency is defined as the ratio of energy output in the form of electrical work divided by the radiant solar energy absorbed. Commercial silicon photocells, which constitute more than 90% of the market, have efficiencies of around 16%. There are, however, many experimental configurations which have much higher efficiencies, as shown in figure 6.11, which gives a list of different types of PV cells under development and the efficiency of laboratory prototypes.

Material with a single band gap can only absorb a small range of photon frequencies with the result that a single junction silicon solar cell with a band gap of 1.34 eV has a theoretical maximum efficiency of 33.7% (Polman *et al* 2016). To improve this, multi-junction cells use a combination of materials with different band gaps such as silicon and perovskite compounds[8]. A larger range of energies can be captured by the different band gaps in separate parts of the cell in this way.

The traditional PV cell is made from a single crystal which is expensive to make because the process must result in a large, pure crystal. For this reason amorphous cells which are cheaper and can be flexible are being investigated, even though their energy efficiency is lower than a solid crystal. Dye and organic cells use photons

---

[8] Pure silicon has a band gap of 1.1 eV corresponding to a wavelength of 1127 nm, and perovskite compounds have band gaps between 1.5 and 2.3 eV corresponding to wavelengths of 826 nm and 539 nm.

which are first absorbed by a dye or organic compound to free electrons which then pass into an n-type semiconductor. The efficiencies are also low but they are very cheap to manufacture compared to single crystal cells. Quantum dot solar cells use nano-size junctions rather than larger p–n crystal junctions. They also have low efficiencies but have the advantage that their band gap can be more easily adjusted in the manufacturing process. Lenses and/or mirrors can be employed to concentrate solar energy on a solar cell which does not increase the energy per area collected but can reduce the cost since the cell itself can be smaller. There are many other configurations being explored, as indicated in figure 6.11.

In the case of PV cells the EROEI is most easily calculated in terms of the energy payback time. It takes about 2.2×10^9 J of energy to make a square meter of single crystal silicon PV cells and 1.6×10^9 J for multi-crystalline silicon. With a cell efficiency of 12% and using 250 W m^{-2} for the incoming radiation it will take

$$(2.2 \times 10^9 \text{ J m}^{-2})/(0.12 \times 250 \text{ W m}^{-2}) = 7.3 \times 10^7 \text{ s} = 2.3 \text{ years} \qquad (6.15)$$

of cell operation to make up the energy used to create a single crystal photocell. PV cells drop in efficiency over time but the projected lifetime for most commercial designs is more than 20 years and in fact the very first photocell ever made can still produce electricity. If the above PV cell lasts 20 years it will produce another 2.0×10^{10} J of energy, a factor of ten times the amount needed to manufacture it. As the technology for making solar cells improves it is expected that energy payback times will drop.

Now that PV solar cells are becoming cost competitive with other energy sources the main disadvantage is the intermittent and disperse nature of sunlight. Storage of electricity is clearly an important problem to solve if solar energy is going to have general utility. Some of the current ideas for large scale energy storage are discussed in the next chapter. Widespread use of solar energy from PV cells also does not fit the current central distribution model for electricity used in most of the world and power companies are struggling with how to integrate distributed, intermittent sources such as solar and wind into the grid.

A final concern is that PV cells create direct current, requiring electronic technology to convert this to the alternating current required in most electronic equipment in use today. This increases the initial cost of implementation and adds a layer of infrastructure needed to use this resource.

6.5.2 Thermal solar

In large scale solar thermal systems the temperature of a working fluid is raised to between 300 °C and 1000 °C using large arrays of mirrors. Two main collection methods involve either a parabolic trough of mirrors focusing energy on a pipe or multiple flat mirrors, motorized to follow the Sun, focusing sunlight on a central tower. Lenses and smaller scale dish reflectors have also been studied. Fluids which have been used to transfer the energy from the collector to a heat engine or other use include water, various gases and molten metals, such as sodium and molten salt compounds. The heated fluid can be used to run a traditional steam engine to

Figure 6.12. Ivanpah solar thermal towers in southern California seen from an airplane. Each of the bright spots is one tower surrounded by 173 500 mirrors which reflect the Sun's energy onto the tower. (Photograph taken by the author.)

generate electricity or the heat can be used directly to run a Stirling engine. The Carnot efficiency for these processes is quite high since temperatures in a solar tower can reach more than 1000 °C. In some cases the thermal heat capacity of the heated fluid is high enough that power can continue to be generated even after the Sun has gone down, providing up to several days of energy storage.

There are more than 60 large scale thermal solar plants operating globally with another 40 either under construction or announced. The total capacity of existing plants is around 4800 MW with 5700 MW under construction. Some idea of the intensity of the collected solar energy is indicated by the areal view of the four towers at the Ivanpah plant in figure 6.12. The costs of building such a system make them economically unattractive without subsidies, but as with all technology, cost is projected to go down as the technology matures.

6.5.3 Thermoclines and salt gradients

The Sun heats the surface of the oceans to above 22 °C in many locations and because of the unusual thermal expansion of water, temperatures at the bottom of the deep ocean are typically 4 °C. This temperature difference can be used to drive a heat engine with a low maximum thermal efficiency of 6% or 7%. Although the efficiency is low, because oceans cover two thirds of the planet there is a tremendous amount of potential energy available, possibly as much as 4500 EJ even at this low efficiency. The largest potential is in the south Pacific where surface temperatures are highest. Ocean thermal energy conversion (OTEC) received a significant amount of interest in the 1970s during the global oil crisis. Unfortunately this energy source is very disperse making it difficult to capture. Although several experimental systems have been built since the 1970s, the commercial prospects of OTEC remain in the future, awaiting better technology.

There have been a number of experiments attempting to enhance the temperature gradient of a body of water by using a salt gradient (Boyle 2004). Visible solar radiation will penetrate to the salty layer at the bottom of an artificial pond but IR radiation from the hot salt solution is trapped by the upper layers of less salty water. This occurs naturally in some salt lakes such as Solar Lake in the Sinai Peninsula of Egypt, where water temperatures reach as high as 60 °C. A 100 °C temperature gradient would yield a 27% efficient Carnot cycle, much higher than possible with OTEC. If a transparent insulating layer is placed on top of a body of water and the body is insulated from heat loss, temperatures can rise high enough to generate steam (Miller 2016).

6.5.4 Solar assisted chemistry

Some chemical reactions require heat to drive them, for example, methane and carbon dioxide will react to form hydrogen and carbon monoxide in the reaction $248\ \text{kJ} + CH_4 + CO_2 \rightarrow 2H_2 + 2CO$, where 248 kJ of heat is applied. Solar energy can be used to drive the reaction and the products can be reacted with catalysts to form methanol, CH_3OH, which is a portable fuel. In some cases the light itself also acts as the catalyst (a photocatalytic process; see Osterloh (2017)). As an alternative, the products can be reacted with steam to generate more hydrogen and CO_2. These and similar reactions provide a way to store solar energy in a liquid fuel for transportation. In a vehicle motor the reaction can be run in reverse to give off the 248 kJ as heat used to do mechanical work.

Artificial photosynthetic processes using photoelectrochemical cells are still in the development stage (Lewis 2016). These systems use sunlight to supply the energy needed for a chemical reaction on a catalytic substrate;—either a semiconductor based electrode or a metal complex. The reaction produces electrons that cause the disassociation of water into hydrogen and oxygen at a second electrode. The process shortcuts the two-step process of using a PV solar cell to generate current for electrolysis by using catalysts to directly disassociate water into oxygen and hydrogen which can be used as a fuel. Similar solar driven chemical processes are being assessed for the potential to turn carbon dioxide and water into fuel stock.

6.6 Tides, waves and other

Several other renewable sources of energy have been investigated with varying degrees of interest. Technical advances may eventually make some or all of these resources important sources of energy but for a variety of reasons none of these resources have yet to be developed on a large commercial scale.

6.6.1 Tides

Based on a calculation of the rate of slowing of the Earth's rotation due to tides there could be as much as 100 EJ of energy in the tides. In some parts of the world, for example the Bay of Fundy in Canada, the tide varies by as much as 17 m (figure 6.13). If water is allowed to flow into a reservoir as the tide rises and then is trapped and forced to flow through turbines to escape as the tide subsides, a system for generating electricity can be fashioned, much like a low dam. The height difference between low and high tide vary greatly around the world and are sometimes no more than half a meter but there are special locations where the resonance frequency of water sloshing back and forth in a bay is close enough to the tidal periodicity to produce large tides. Other locations besides the bay of Fundy with large tides are the Severn Estuary in the United Kingdom, Solway Firth in the UK and the Gulf of Khambhat in India. A number of experimental tidal generation plants have been built, with the largest commercial tidal plant built to date at the La Rance Tidal Basin at Saint Malo, France, which generates 0.002 EJ of electricity a year from tides that vary by about seven meters.

Damming tidal basins has some of the same environmental impacts as building large dams, including significant environmental impact on local estuaries on which a substantial portion of the world's population depends for food. In consideration of these issues many plans for capturing tidal energy involve submerged turbines operating on the flow of water rather than a hydraulic head captured by a dam.

It is difficult to calculate the percentage of the total tidal energy that can be captured except for the few places where dams can be constructed. For submerged

Figure 6.13. Low tide in the Bay of Fundy. At high tide the water reaches the tree line. (Photograph taken by the author.)

turbines the amount of energy that can be extracted is highly dependent on the local site since narrow channels can increase water flow speed substantially. One estimate of energy available from the best 28 sites in the world is about 10 EJ (Smil 2005).

6.6.2 Waves

A sinusoidal wave carries an average energy per unit surface area of

$$\frac{E_{\text{ave}}}{\text{area}} = \rho g y^2/2, \tag{6.16}$$

where ρ is the density of the liquid, g is the acceleration of gravity and y is the amplitude of the wave. Multiplying by $v = f\lambda$ gives the power per meter width of the wave as

$$\frac{P_{\text{ave}}}{l} = \rho g v y^2/2. \tag{6.17}$$

If we assume an average height of one meter for ocean waves near the coast, a frequency of about 0.1 Hz, wavelengths of 10 m and speeds of 1 m s^{-1} we obtain a power of $\frac{P_{\text{ave}}}{l} = 5\ \text{kW m}^{-1}$, five kilowatts per meter of coastline. The US has an eastern coastline of 3300 km so the power available at the East Coast is on the order of $5\ \text{kW m}^{-1} \times 3300\ \text{km} = 1.65 \times 10^{10}$ W. Assuming this constant rate of energy flow, about 0.5 EJ could be captured in a year if the capturing process is 100% efficient. For all of the coast of the US, including Alaska, the figure is about 7.5 EJ. The amount of energy available in the open ocean is, of course, much greater than what can be captured at shore but, as was the case with offshore wind and OTEC resources, concentrating and transporting the energy from where it is captured to where it can be useful is problematic.

A surprising number of devices have been tested to capture wave energy and an even larger number of designs have been proposed (Boyle 2004, WEC 2013). Two devices that have actually be constructed, the Pelamis 'sea snake' and Salter's Duck, use the rocking motion of the waves to pressurize hydraulic fluid which can then be used to turn a generator for electric energy. There are many other designs using floats attached to stationary anchors so that the up and down or side to side motion of the waves moves levers to capture kinetic energy. Overtopping or surge mechanisms operate closer to the shore and trap water splashing over a barrier to form a hydraulic head similar to tidal energy capture. The oscillating motion of the ocean near shore can also be allowed to force water through a one-way valve to compress a working fluid which then provides mechanical energy to a generating system.

The Pelamis system was the first large scale device to successfully capture significant amounts of wave energy and connect that energy to the electric grid in Orkney, Scotland. Several other devices were placed off the coast of Portugal in 2008 but were later shut down due to high seas which broke the systems from their moorings, and the Scottish company making the devices has now gone out of business. Given the technical difficulty it is not clear when wave energy will become a viable energy source.

6.6.3 Other

Humans are creative and countless other proposals for tapping renewable energy flows around the globe have been made. Various schemes have involved using thermal energy from underground nuclear explosions, the capture of high speed winds in the upper atmosphere with kites or balloons with turbines attached, solar collecting satellites above the Earth's atmosphere which would beam energy to the Earth's surface using microwaves and the possibility of collecting the energy in lightning.

The energy from mixing when freshwater from rivers mixes with the ocean is another possible energy source. If two liquids with different salinities are separated by a semi-permeable membrane that allows water through but not ions, a pressure difference will build up between the two liquids. This osmotic pressure difference can be used to drive mechanical systems such as electric generators. The estimated energy per liter for mixing salt water with fresh water is 2.2 kJ per liter of fresh water. If the global discharge of fresh water into the oceans is around 45 000 km^3 per year this represents an energy resource of about 10 EJ, although it is not clear how efficiently this energy could be extracted.

The electrochemical potential difference between salt and fresh water can also be used to generate a voltage source in a process called reverse electrodialysis. In this process a series of positive and negative ion exchange membranes are inserted between salt and fresh water which will separate the sodium and chlorine atoms forming a potential difference between the ends of the membrane sequence (IPCC 2012).

6.7 Potential and cost of renewables

Estimates of how much renewable energy is available vary widely and are highly subject to the assumptions made about the technology available for capturing the energy (table 6.2). For example, including energy captured far from shore over the open ocean, an uncertain prospect, boosts the potential of wind, wave and solar by two orders of magnitude. There is some general agreement among most analysts as to the general order of magnitude of renewable sources, however. Most references concede that solar resources are by far the most abundant, probably far exceeding global demand use for the next century, even assuming no technological advances. Wind resources are typically estimated to be two to three orders of magnitude less than solar but still larger than current global demand. Recoverable biomass appears to also be large enough to play a part, perhaps as much as half of the global energy demand. The technology for capturing wave and tidal energy is still under development so it is not clear how much energy these sources might be able to supply in the near future, most estimates are two orders of magnitude less than wind resources. Geothermal will likely play a small role except in locations where there are exceptional geothermal resources.

Given the vast renewable resource available from wind and solar, one might be puzzled as to why a higher percentage of the world's energy does not come from renewables. The answer to the question of why the world still provides 80% of its

Table 6.2. Very rough estimates of annual global renewable energy resources. The figures are based on the following sources: (WEC 2013, Smil 2005, Sorensen 2000, Boyle 2004, Hafemeister 2007, Moriarty and Honnery 2012, EIA 2016, IPCC 2012 and REN21 2016).

Source	Estimated total energy (EJ)	Estimated recoverable (EJ)	Current use (EJ)
Solar	1×10^6 over land	100–30 000	2.2 thermal, 0.9 PV
Wind	11 000 over land	50–1000	3
Biomass	2850	12–1200	50
Hydroelectric	300–900	45–60	37
Tidal	90	10	0.01
Wave	60	15–30	0.01
Geothermal[a]	13 000	100	1

[a] Estimates vary depending on location and depth. The difference in available energy between a depth of 3 km and 10 km is a factor of 10.

energy with fossil fuels is an economic one. Table 6.3 shows the various estimated costs involved in different future sources of electricity, anticipated by the EIA to be available in 2018. The numbers shown are the costs in dollars per MWh of electricity generated in today's dollars, assuming the plant construction uses current technology. Lead times vary depending on the source, with nuclear plants taking six years or more to construct and wind and solar plants on the order of three years. Likewise, most power plants are built to last 50–60 years and it is much cheaper to continue generating power with an existing plant than to shut that plant down prematurely and build a new, albeit more efficient plant[9]. A state by state plan outlining the technical details and costs of converting all energy use in the US to renewables by 2050 using current technology is given in Jacobsen *et al* (2015) and similar plans for 139 countries around the globe are available in Jacobsen *et al* (2016).

The capacity factor indicates what percentage of the time the source is expected to be online. Nuclear power plants seldom have to be shut down for maintenance and refueling so they have high capacity factors of 90%. The low capacity factors for wind and solar reflect the uncertain and intermittent nature of these sources. Gas turbines are typically used to supply energy when demand is high and are not expected to be online all the time. Levelized capital cost is the cost to build the plant, amortized over the life of the plant (see chapter 10 for comments about how this is calculated). Fixed O&M are normal operation and maintenance costs whereas variable O&M includes fuel and unexpected maintenance. Transmission costs are higher for solar and wind because it is expected that these sources will be located at a distance from where the electricity will be used.

Most estimated future costs for fossil fuels include the cost of *carbon sequestration*; capturing the carbon dioxide being emitted. The most investigated technique for separating carbon dioxide from flue exhaust is scrubbing with amine compounds

---

[9] In economics this is referred to as marginal product.

Table 6.3. US projected costs for 2018 in 2015 dollars for a MWh of electricity from various sources. CCS is carbon sequestration, now required for new coal power plants in the US. CC is combined cycle, where exhaust heat from one generating stage is used to generate more electricity (EIA 2016).

	Capacity factor (%)	Levelized capital cost	Fixed O&M	Variable O&M, fuel	Transmission	Total	Tax credit	Total with credit
Advanced coal with CCS	85	97.2	9.2	31.9	1.2	139.5	N/A	139.5
Conventional CC gas	87	13.9	1.4	41.5	1.2	58.1	N/A	58.1
Advanced CC gas	87	15.8	1.3	38.9	1.2	57.2	N/A	57.2
Advanced CC with CCS gas	87	29.2	4.3	50.1	1.2	84.8	N/A	84.8
Conventional combustion turbine gas	30	40.9	6.5	59.9	3.4	110.8	N/A	110.8
Advanced combustion turbine gas	30	25.8	2.5	63	3.4	94.7	N/A	94.7
Advanced nuclear	90	78	12.4	11.3	1.1	102.8	N/A	102.8
Geothermal	91	30.9	12.6	0	1.4	45	−3.1	41.9
Biomass	83	44.9	14.9	35	1.2	96.1	N/A	96.1
Wind	40	48.5	13.2	0	2.8	64.5	−7.6	56.9
Wind, offshore	45	134	19.3	0	4.8	158.1	−11.4	146.7
Solar PV	25	70.7	9.9	0	4.1	84.7	−18.4	66.3
Solar thermal	20	186.6	43.3	0	6	235.9	−56	179.9
Hydro-electric	58	57.5	3.6	4.9	1.9	67.8	N/A	67.8

which bind CO_2 at low temperatures but will release it at high temperatures. Other techniques such as adsorption by metal compounds and membranes are under investigation. Once captured from flue exhausts the CO_2 can be compressed and injected into deep (two or more kilometers) underground geological formations such as salt domes and depleted oil or gas wells. The carbon dioxide can also be used in secondary oil extraction to push oil out of reservoirs where natural pressure is not sufficient. There are uncertainties about the stability of long term underground storage of CO_2, but dozens of large scale sequestration projects are underway around the world.

It is instructive to compare the numbers in table 6.3 to similar projections made by the US EIA in 2012. In that year, future solar PV total cost was estimated to be $360/MWh, more than three times the cost of any fossil fuel. In the same year, however, onshore wind was $91/MWh, very comparable to fossil fuel costs. These figures were fairly predictive; the competitive cost of wind has made wind energy the fastest growing renewable energy resource for over a decade. Current projections have solar PV, even without subsidies, in a cost competitive regime with fossil fuels so we can expect to see a sharp increase in solar energy installations in the next few years, matching the recent steep increase in wind energy. It is also the case that energy prices vary significantly with location. Currently natural gas and wind are cheaper energy sources than coal in most locations in the US (Energy Institute 2017).

Fluctuating fuel cost is another factor which has a large impact on which energy source is most likely to be pursued. The price of oil has fluctuated from $30 a barrel in the 1990s to a high of $150 a barrel in 2008[10]. Presently the price of oil is back down to just under $50 a barrel. At $150 a barrel, all other energy sources, including offshore wind are cheaper than oil. At $30 a barrel onshore wind energy is twice as expensive as petroleum. One of the reasons most new power plants being built today are fueled with natural gas is that hydraulic fracturing has brought the cost of gas down by a factor of six from a high in 2008 of $12 per million BTUs. These prices are expected to remain low for the near future.

One number, which is not shown in the table but should be kept in mind for comparison purposes is the price of conservation measures. It is estimated that about $60 of conservation efforts, for example more efficient heating and cooling for buildings, will save a MWh of electricity. Improving energy efficiency is the cheapest and most effective way to reduce energy costs and reduce carbon emissions. Many more details can be found in (ACEEE 2017).

Projects

1. Write a report summarizing what is known about environmental problems associated with large dam projects. Include a discussion of ecological issues, health and population problems, carbon emissions, political issues, and geological problems such as local earthquakes.

---

[10] Oil averaged around $20/barrel during the 1950s, 1960s and early 1970s.

2. Write a report outlining the conflict between experts who think bio-fuels will not impact food supplies and those who do. Include as many statistics and data as are available. What is the current status of this debate?
3. Investigate what is known about the carbon balance of growing plants for fuel. Attempt to answer the question 'Are bio-fuels carbon neutral?' including any applicable data you find.
4. Find the wind resource map for a particular region of the globe and calculate the possibility of supplying all the energy used in this region. A useful source is the Solar and Wind Energy Resource Assessment (SWERA) from the UN http://en.openei.org/apps/SWERA/. Write a review of your findings.
5. Visit Earth https://earth.nullschool.net/. Analyze and write a review of the global visualizations from computer simulations of ocean currents, wind speeds at different heights, relative humidity and particulates in the atmosphere. Comment on potential locations of renewable energy development projects using wind and ocean currents. Attempt to verify your observations with other data.
6. Write a report that analyzes and compares the major tidal energy projects in Shiwa Lake, South Korea; La Rance, France; Swansea Bay, UK; MayGlen, Scotland; Annapolis, Canada; and any others that you can find. What lessons have been learned from these projects that might be applicable to future renewable energy sources?
7. Find global maps of solar energy resources and comment on where the resources are located. Write a report on current plans to tap solar energy from the Sahara desert on a large scale for use in Europe. What other large scale plans are under development for solar energy?
8. Expand table 6.2 to include several other countries. Analyze the assumptions being made and estimate the range of possible errors for these assumptions. How sensitive are the results on these assumptions?
9. Review the calculations given in the supplemental material in Jacobsen *et al* (2015) to verify the claims made in the body of the paper that the US could convert completely to renewable energy by 2050 using current technology (http://web.stanford.edu/group/efmh/jacobson/Articles/I/WWS-50-USState-plans.html).
10. Summarize and critique a section of the IPCC report 'Renewable energy sources and climate change mitigation' (IPCC 2012).
11. Write a report on the current status of any of the solar assisted chemistry ideas mentioned in the text.
12. Calculate the energy available per year from surface lightning strikes. Should this idea be pursued?

References

ACEEE (American Council for an Energy-Efficient Economy) 2017 http://aceee.org/ (Accessed: 29 January 2017)

Boyle G 2004 *Renewable Energy; Power for a Sustainable Future* (Oxford: Oxford University Press)

BP 2016 Statistical Review of World Energy www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html (Accessed: 7 August 2016)

Denker J S 2014 See How it Flies, A new spin on the perceptions, procedures, and principles of flight www.av8n.com/how/#mytoc Accessed: 17 August 2016

EIA (US Energy Information Administration) 2016 Annual Energy Outlook www.eia.gov/forecasts/aeo/index.cfm and www.eia.gov/forecasts/ieo/ieo_tables.cfm (Accessed: 17 August 2016)

Energy Institute, University of Texas, Austin 2017 Gull levelized cost of electricity in the United States by county http://calculators.energy.utexas.edu/lcoe_map/#/county/tech (Accessed: 2 January 2017)

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Hafemeister D 2007 *Physics of Societal Issues* (New York: Springer)

ICLD (International Commission on Large Dams) 2016 www.icold-cigb.org/GB/World_register/general_synthesis.asp (Accessed: 9 September 2016)

IEA (International Energy Agency) 2016 Renewables information www.iea.org/topics/renewables/ (Accessed: 6 January 2017)

IHA (International Hydropower Association) 2016 Hydropower Status Report www.hydropower.org/ (Accessed: 17 August 2016)

IPCC (International Panel of Climate Change) 2012 Renewable Energy Storage and Climate Change Mitigation www.ipcc.ch/pdf/special-reports/srren/SRREN_Full_Report.pdf (Accessed: 6 January 2017)

Jacobsen M Z, Delucchi M A, Bazouin G, Bauer Z A F, Heavey C C, Fisher E, Morris S B, Piekutowski D J Y, Vencill T A and Yeskoo T W 2015 100% clean and renewable wind, water, and sunlight (WWS) all-sector energy roadmaps for the 50 United States *Energy Env. Sci* **8** 2093

Jacobsen M Z, Delucchi M A, Bazouin G, Bauer Z A F, Goodman S C, Chapman W E and Cameron M A *et al* 2016 100% clean and renewable wind, water, and sunlight (WWS) all-sector energy roadmaps for 139 countries of the world http://web.stanford.edu/group/efmh/jacobson/Articles/I/CountriesWWS.pdf (Accessed: 5 January 2017)

Korner C 2017 A matter of tree longevity *Science* **355** 6321 p 130

Kromdijk J, Głowacka K, Leonelli L, Gabilly S T, Iwai M, Niyogi K K and Long S P 2016 Improving photosynthesis and crop productivity by accelerating recovery from photoprotection *Science* **354** 857

Kumar A *et al* 2011 *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation* ed O Edenhofer *et al* (Cambridge: Cambridge University Press)

Lewis N S 2016 Research opportunities to advance solar energy utilization *Science* **351** 354

Miller J 2016 Solar steam generator needs no lenses or mirrors *Phys. Today* **69** 17

Moriarty P and Honnery D 2012 What is the global potential for renewable energy? *Renew. Sust. Energy Rev.* **16** 244–52

NRCS (Natural Resources Conservation Service) www.wcc.nrcs.usda.gov/climate/windrose.html (Accessed: 17 August 2016)

NREL (National Renewable Energy Laboratory) 2016a www.nrel.gov/wind/data-tools.html (Accessed: 14 November 2016)

NREL (National Renewable Energy Laboratory) 2016b www.nrel.gov/pv/national-center-for-photovoltaics.html (Accessed: 14 November 2016)

NREL (National Renewable Energy Laboratory) 2017 dynamic maps, GIS data, and analysis tools http://www.nrel.gov/gis/solar.html (Accessed: 4 April 2017)

Osterloh F E 2017 Photocatalysis versus photosynthesis: a sensitivity analysis of devices for solar energy conversion and chemical transformations *ACS Energy Lett.* **2** 445

Polman A, Knight M, Garnett E C, Erhler B and Sinke W C 2016 Photovoltaic materials: present efficiencies and future challenges *Science* **352** doi:10.1126/science.aad4424

REN21 (Renewable Energy Policy Network for the 21st Century) 2016 RENEWABLES 2016 Global Status Report www.ren21.net/status-of-renewables/global-status-report/ (Accessed: 6 January 2017)

Smil V 2005 *Energy at the Crossroads* (Cambridge, MA: MIT Press)

Smil V 2015 *Harvesting the Biosphere: What We have Taken from Nature* (Boston, MA: MIT press)

Sorensen B 2000 *Renewable Energy: Its Physics, Engineering, Environmental Impacts, Economics and Planning* (New York: Academic)

Sovacool B K 2012 The avian and wildlife costs of fossil fuels and nuclear power *J. Integrative Env. Sci.* **9** 255–78

SWERA (Solar and Wind Energy Resource Assessment) On-line data tool http://en.openei.org/wiki/SWERA (Accessed: 6 January 2017)

WEC (World Energy Council) 2013 World Energy Resources 2013 Survey www.worldenergy.org/ (Accessed: 17 August 2016)

Weisz P B 2004 Basic choices and constraints on long-term energy supplies *Phys. Today* **57** 47

Zhu X, Long S and Ort D 2008 What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? *Curr. Opin. Biotechnol.* **19** 153

IOP Concise Physics

Physics and the Environment

Kyle Forinash III

Chapter 7

Energy storage

Several aspects of energy storage are discussed in this chapter. In general the most important factors when considering an energy storage application are *energy density*, which is the energy per mass of a source, and *power density*, which is the rate at which the energy can be provided per mass (also referred to as specific power). Energy storage for transportation (fuel) must be lightweight with high energy density and be able to supply energy rapidly for acceleration. On the other hand, large scale energy storage for load leveling of the electric grid or for renewable energy is not restricted by weight. We start with embodied energy, the energy contained in a material or device, including the energy needed to make it.

7.1 Embodied energy

One interesting type of stored energy that is often overlooked is embodied energy (Ashby 2013). Embodied energy is the energy involved in manufacturing and using a particular appliance or other modern accessory, including the energy needed to extract material of which the object is made. For example most of the energy used in the normal life cycle of a cell phone is consumed in its manufacturing (420 MJ versus 85 MJ in power consumption)[1]. In comparison automobiles use almost five times the amount of energy used as fuel (1.9 GJ) in their normal life time as the amount of energy it takes to make the vehicle (0.4 GJ)[2]. To reduce energy consumption due to cell phone use requires changes to its manufacturing process but for cars, altering fuel efficiency would have a greater impact on total energy consumed (Ashby 2013). Plastics of all types have a very high embodied energy, comparable to gasoline, which makes recycling and modern landfills suitable targets for recapturing energy from waste. Embodied energy also includes the energy cost for obtaining or making

---

[1] According to the Consumer Technology Association, average cell phone lifespans are around four years.

[2] The average age of cars on the road in the US has risen from eight years before 2000 to around 11 years currently.

doi:10.1088/978-1-6817-4493-3ch7

Table 7.1. Energy cost of various common construction materials (Smil 1999).

Material	Energy cost (MJ kg^{-1})	Source
Aluminum	227–342	Bauxite
Bricks	2–5	Clay
Cement	5–9	Clay and limestone
Copper	60–125	Sulfide ore
Glass	18–35	Sand
Iron	20–25	Iron ore
Limestone	0.07–0.1	Sedimentary rock
Paper	25–50	Timber
Sand	0.08–0.1	Riverbed
Wood	3–5	Timber

a particular material or device. The energies required to obtain a few basic construction materials are shown in table 7.1.

The energy gain from recycling compared to refining a substance from raw sources depends on the material. Producing a kilogram of aluminum from ore takes about 300 MJ of energy but it takes less than 20 MJ to produce a kilogram of aluminum from recycled scrap, which is an energy savings of 93%. Similarly for copper there is about a 75% saving when using scrap copper compared to mining and refining copper ore. Manufacturing a kilogram of iron from ore takes about 25 MJ but only 1.3 MJ is needed to produce iron from scrap iron and using recycled paper as stock for a paper mill results in about a 40% saving in energy. On the other hand, recycling glass containers results in less than a 30% energy saving over manufacturing glass directly from unprocessed sources (The Economist 2007).

Recycling plastic is only energetically favorable if the many different types of plastics, shown in table 7.2, can be separated from each other. Melting and ignition temperatures are different for each type, which means that the recycling process will be ruined if a batch of PET bottles is contaminated with a bottle made with PVC, for example (Hummel 2006). If the plastics are properly separated there can be as much as a 70% energy saving over making the plastic directly from petroleum products but separation is labor intensive which makes it costly.

7.2 Transportable fuel

The energy content of several common substances, some of which are in use today as fuels, is shown in table 7.3. As we saw in chapters 2 and 3, gasoline and diesel engines are not particularly efficient in converting primary energy into useful mechanical energy due to constraints imposed by the second law of thermodynamics. It is only the high energy per kilogram available from petroleum distillates that makes those types of motors effective. If it were not for considerations of carbon emissions and global petroleum reserves there would be no motivation for switching away from petroleum based fuels for transportation. Burning petroleum,

Table 7.2. Types of plastics and their Resin Identification Codes. Melting points depend on the specific application (Hummel 2006).

Material	Use	Abbreviation	Melting temperature	Resin ID code
Polyethylene terephthalate	Beverage bottles, cups, packing	PET	250 °C	1
High-density polyethylene	Bottles, cups, milk jugs	HDPE	120 °C–180 °C	2
Low-density polyethylene	Plumbing, siding, flooring	PVC	100 °C–200 °C	3
Low-density polyethylene	Bags, packaging	LDPE	95 °C	4
Polypropylene	Food containers, fiber	PP	130 °C–171 °C	5
Polystyrene	Plastic utensils, Styrofoam	PS	240 °C	6
Other (acrylic, nylon, polycarbonate)	Fiber, CDs, DVDs, coatings	PLA	190 °C–350 °C	7

however, produces carbon dioxide and other pollutants which, as we will see in chapter 9, constitute a significant global climate problem.

Nuclear fuel sources offer enormous energy per kilogram but are unsuitable for transportation for safety reasons. The amount of shielding needed to protect passengers and the surroundings from radiation and to prevent a spill in the case of an accident effectively relegates the role of uranium as a fuel to stationary power plants.

Hydrogen is attractive as a potential fuel due to its high energy content per weight and the fact that it produces only water when burned. But hydrogen is difficult to contain (and is in fact often used as a leak detector in high vacuum systems). Current technology requires a sturdy storage tank to contain liquid or compressed hydrogen, which makes the energy to weight ratio much less attractive. Future technology, for example nano-tubes or gas–solid adhesion may some day improve the utility of hydrogen as a fuel. It is also the case that, unlike fossil fuels, it is not found in a pure state in nature and requires energy to manufacture[3]. Other fuel in gaseous form such as propane, natural gas and compressed air face similar problems, but these may not be insurmountable and in fact vehicles using these fuels are already available in the market.

Although the earliest cars were electric the low energy density of lead acid batteries eventually resulted in a switch to gasoline based motors for transportation. Lithium batteries contain five times the energy per kilogram compared to lead batteries and all-electric cars exist using these energy storage devices, but ranges are low compared to gasoline powered cars and the recharging times inhibit long distance travel. A more detailed discussion of batteries is found below.

---

[3] As we will see in chapter 8, commercial hydrogen today is mostly derived from steam reforming of fossil fuels.

Table 7.3. Energy content per weight and per volume for various substances (Smil 1999).

Fuel	Energy content MJ kg^{-1}	Energy density MJ L^{-1}
Natural uranium (0.7% U-235)[a]	159 000	3×10^6
Hydrogen (STP)	114.0	10
Hydrogen (liquid, with tank)	10	10
Hydrogen (compressed gas, with tank)	5	5.6
LPG propane	49.6	25.3
Gasoline	47.9	32.1
Diesel	45.8	38.7
Plastic (polyethylene)	42.2	42.6
Crude oil	42.0–43.0	37
Plant oil (bio-diesel)	38–42	~33
Butter	30	27
Ethanol	29.6	22
Coal	22–32	~72
Natural gas (STP)	20.0	0.03
Cereal grains, bread	~15	~26
Air dried wood	12–15	2–3
Straw, dung	12–15	2–3
Meat	5–10	5–10
Potatoes	4	7.4
Beer	1.8	1.8
Carbon fiber flywheel	0.8	0.1
Fruit and vegetables	0.6–1.8	0.6–1.8
Lithium batteries	0.5–2.5	1.0
Compressed air, with tank	0.1	0.1
Lead batteries	0.1	0.15

[a] Based on the amount of electricity which can be generated at 30% efficiency according to the US Department of Energy. The energy density of pure U-235, calculated from the energy released per fission per molecule is 79 390 000 MJ kg^{-1}.

The values in table 7.3 also explain the current interest in bio-diesel and ethanol as a fuel for vehicles, although as we shall see in the next chapter, there are problems with these fuels as well. It would be possible to fuel automobiles with wood, straw or vegetables, but the energy per weight makes them unlikely fuel sources for transportation. Coal, with half the energy per kilogram as gasoline, was once used as a fuel for steam locomotives but with the advent of diesel fuel, a lighter weight alternative, coal fired trains have mostly been abandoned except as novelty items and in a few places where there is no alternative.

Diet plans such as Weight Watchers generally recommend eating fruits and vegetables as opposed to meat and potatoes and the reason should be clear from the data in table 7.3. A kilogram of fruits and vegetables contain half the calories so it is possible to fill up without taking on excess energy to be stored as fat. For workers

Table 7.4. Energy density versus power density for select fuels and mechanisms (Smil 2005).

Source	Energy density MJ kg^{-1}	Power density W kg^{-1}
Super-capacitors	0.05	5000
Flywheels (conventional)	0.05	400
Flywheels (advanced)	0.1	3000
Lithium ion battery	0.3	350
Ni/Zn battery	0.2	300
Lead battery	0.1	70
Ethanol	29.6	100
Gasoline	47.0	100
H_2 fuel cell	10.0	50

doing physical labor and athletes, however, carbohydrate loading with pasta and cereal grains provide calories needed to function well.

The power density of an energy storage fuel or mechanism must also be matched to the application. Gasoline engines in automobiles do not provide as much power at low rotation speed compared to electric motors because the rate at which energy can be provided is lower. The power needed to overcome friction and drag forces for an automobile traveling at slow constant speeds is on the order of 8000 W (approximately 10 hp) for an average car but the power needed for reasonable acceleration is approximately twenty times higher, as shown in the next chapter. Hybrid cars such as the Chevrolet Volt, Ford Fusion and Toyota Prius couple an electric motor with a gasoline engine in an attempt to couple the high energy density of gasoline engine systems with the high power density of electric motors and batteries. Table 7.4 compares energy density with power density for several mechanisms and fuels, data that are often provided in a *Ragone* plot.

The overall energy need to build the energy storage device or collect and refine the fuel must also be taken into account, as well as the emissions and waste involved in the manufacturing and the use of these storage mechanisms. These issues, referred to as *well-to-wheels efficiency* or, if the energy to manufacture and dispose of the vehicle is included, the *life-cycle efficiency*, will be taken up for transportation fuels in chapter 8.

7.3 The need for large scale energy storage

As we saw in chapter 6, many renewable energy sources have low capacity factors meaning that they are intermittent supplies. Solar and wind energy, the fastest growing renewables, in particular are not available all the time and so will require energy storage on a large scale. Some of the technology for providing uninterrupted electricity has already been developed and is in use, albeit on short time scales. The electric grid has to respond to demand on a second to second time scale requiring short term load leveling energy storage systems. There are also daily and seasonal changes in electricity demand that require energy storage on medium and longer time scales. Examples of energy storage systems for these applications include

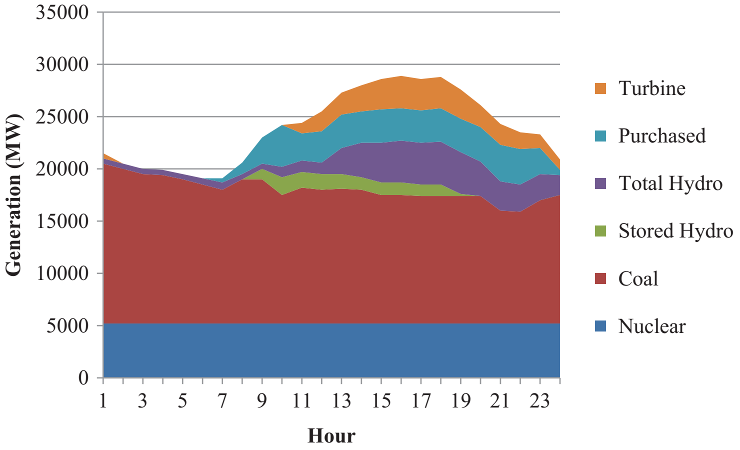


Figure 7.1. An example of daily electrical consumption by source for a 24 h period based on figures from the Tennessee Valley Authority for the Raccoon Mountain service region. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

pumped hydroelectric, flywheels, thermal storage, batteries, flow batteries and super-capacitors as considered in the following sections.

As can be seen in figure 7.1, energy demand changes over a 24 h period and is typically supplied by different sources. Nuclear energy, with a high capacity factor, supplies a base load amount of electricity. Electricity from coal is the next cheapest source and is used to supply the bulk of the electricity. Natural gas turbines can be brought on line quickly to supply peak demand. Hydroelectric is actually the cheapest method for making electricity but because it can respond very quickly to changes in demand it is often reserved for use at peak demand times.

Hydroelectric plants in addition to providing energy can also store energy on a large scale. Water is pumped back into the lake behind the dam when demand is low (typically after 22:00 and before 7:00 AM) to be re-released to provide electricity at peak demand times. This process is not 100% efficient so pumped hydro is a net energy loss, but it does provides additional energy when it is needed during the day and can obviate the need to build additional power plants. This technique can be used even where the geographical locale is flat by using subterranean reservoirs. A cavern with a volume of $4.0 \times 10^6\ \mathrm{m}^3$ at a depth of 2 km can store $U = mgh = \rho Vgh = 7.8 \times 10^{13}$ J of energy. Over a 10 h period this can supply $U/t = 2.1$ GW, equivalent to a moderate size generating plant. The energy is extracted in a manner similar to hydroelectric plants; as the water descends from a surface holding tank the kinetic and potential energy is captured by turbines.

One of the more important features of electricity is that it can be transferred relatively rapidly from one region to another. As figure 7.1 shows, a significant amount of electricity for this region is purchased from surrounding suppliers. The redistribution of electricity of a grid is at the same time convenient and also problematic. Several large scale power outages have occurred when parts of an electrical grid fail due to lightning or other causes and the system tries to redistribute too much energy causing a domino effect, where large sections of the grid shut down to protect the wires from currents that are too high[4].

For daily load leveling of electrical demand pumped hydroelectric is the most common form, but other alternatives are used on shorter time scales, for example flywheels. Table 7.5 compares several energy storage devices in use or under development with the energy content of gasoline.

None of the storage mechanisms listed in table 7.5 can match the energy stored per kilogram in gasoline. However for many applications the available power, cycle life and cycle efficiency are much more significant than energy density. *Cycle efficiency* gives the percentage of primary energy lost in storing and retrieving energy from the storage system, for example, pumped hydro typically loses 25%–35% of the energy stored. *Cycle life* is the number of times the system can be recharged before it deteriorates to the point of not being useful as a storage system. Cycle life is important in many applications, such as load balancing in the power grid where energy is constantly being stored and retrieved to keep the flow of electricity constant.

As figure 7.2 shows, the electrical energy use in the US fluctuates significantly over the seasons, primarily due to the increased use of air conditioning during the summer. As seen in the figure, electricity generated from natural gas has overtaken coal as the result of hydraulic fracturing which has made natural gas less expensive. Wind and solar energy have begun to have an impact over the past decade. However at the present time load leveling on a seasonal scale is not being practiced and the net amount of energy storage capacity for the electric grid is small. In the US pumped hydro accounts for 95% of the storage capacity and can provide as much as 24.5 GW. Compressed air capacity is 423 MW, thermal is 431 MW, flywheels 40 MW and batteries 304 MW. This represents about a quarter of the total electrical generating capacity of 1000 GW in the US, however, this output cannot be maintained for long periods of time, typically only for a 24 h cycle. It can be expected that current widespread use of load leveling will increase as intermittent renewable sources are added to the grid, but this will require significant investment in infrastructure.

7.4 Solar thermal energy storage

As we saw in figure 6.8, the picture from Mesa Verde, thermal storage of energy from the Sun has a long history. Applications include solar heating of rocks and slabs of concrete for warmth at night, thick adobe walls to help equalize daytime and nighttime temperatures, as well as water tanks painted black to absorb energy from the Sun during the day, used extensively in warm climates. As we saw in section 6.5 on solar energy, molten salt compounds and metals can be used to transfer and store heat in large solar thermal arrays. Enough energy can be stored in these compounds

---

[4] The Northeast Blackout of 1965 in the US and the Great Storm of 1987 in Europe are two of many such events.

Table 7.5. Comparison of energy storage mechanisms with the energy content of gasoline (Smil 2005, Sorensen 2000, Tester *et al* 2005). MWe stands for equivalent mega-watt electrical energy produced.

Storage type	Energy (kJ kg^{-1})	Potential size (MWe)	Cycle efficiency (%)	Power (W kg^{-1})	Cycle life
Gasoline	47 000				
Pumped hydroelectric	1	100–1000	65–80		50 000
Compressed air	100	50–1000	40–50		10 000
Flywheels		1–10	95		50 000
Steel	30–120				
Carbon fiber	~200				
Thermal					infinite
Water (100 °C–40 °C)	259				
Rocks (100 °C–40 °C)	40–50				
Rocks (400 °C–200 °C)	160				
Iron (400 °C–200 °C)	30				
Salt compounds	300				
Batteries					
Lithium ion	~580	0.5–50	99	350	1200
Nickel cadmium	~350		70–90	150	1500
Nickel hydride	~290	0.5–50	66	250–1000	1000
Lead storage	~140	0.5–100	70–80	70	500–800
Superconducting magnet	100–10 000	10–10 000	85		infinite
Super-capacitors	18–36	1–10	95		100 000
Ice (latent heat of fusion)	334		100%		infinite

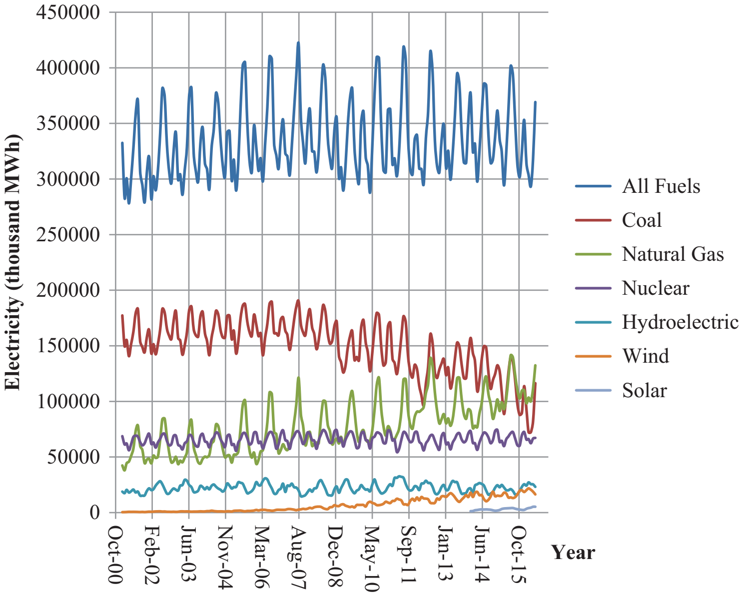


Figure 7.2. Electrical energy consumption by month and by source for the US (EIA 2016).

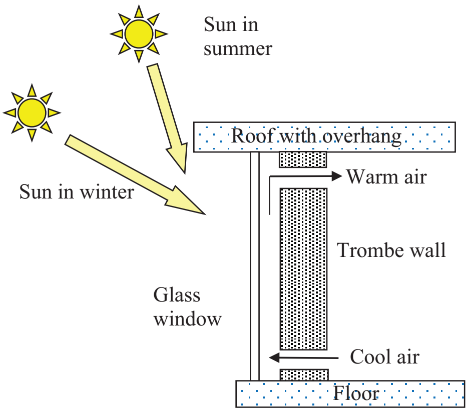


Figure 7.3. Representation of a Trombe wall. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

to run steam generation of electricity at night, after the Sun has set. As an example of unwanted heat storage, the average nighttime temperature of some cities is as much as 3 °C higher than it would be otherwise, in part due to the heat reservoirs created by large amounts of concrete found in cities.

Various forms of the Trombe wall, a schematic of which is shown in figure 7.3, have been in use for hundreds of years. In the winter months solar energy from the Sun is at a lower angle on the horizon and can reach a large thermal mass, chosen to have a high heat capacity. Air circulates past the wall at night, either by natural convection or forced using a fan, moving the heat from the wall into the living area.

A roof with a large overhang prevents the Sun from shining directly onto the wall in the summer when the Sun is higher in the sky.

The energy that can be stored in a Trombe wall is easily estimated from the physical properties of concrete and the approximate average solar flux of 250 W m^{-2} given in chapter 4. Suppose we have a concrete wall 2 m × 5 m and 20 cm thick with density 2400 kg m^{-3} and specific heat capacity of 1000 J $(kg\ K)^{-1}$. It is painted black and assumed to be 50% efficient in collecting energy (the net albedo is 0.5). With these figures we can store $0.5 \times 250\ \mathrm{W\ m^{-2}} \times 2\ \mathrm{m} \times 5\ \mathrm{m} \times 8\ \mathrm{h} \times 3600\ \mathrm{s} = 3.60 \times 10^7\ \mathrm{J}$, about 34 000 BTU which is comparable to a furnace of 10 000 BTU recommended for 180 sq ft (16.7 m^2) of living space in the US. The temperature increase of the concrete during the day is $\Delta T = Q/mC = 3.60 \times 10^7\ \mathrm{J}/(2400\ \mathrm{kg\ m^{-3}} \times 2\ \mathrm{m} \times 5\ \mathrm{m} \times 0.20\ \mathrm{m} \times 1000\ \mathrm{J\ (kg\ K)^{-1}}) = 7.5\ °\mathrm{C}$ for a wall 0.2 m thick of concrete with a density of 2400 kg m^{-3} assuming it does not lose any heat.

We can also find the rate of cooling during the night from this wall using Newton's law of cooling for the case where the cooling object's temperature is allowed to change over time[5]. We have $T(t) = T_S + (T_0 - T_S)e^{-(\lambda t)}$ where the initial temperature of the cooling wall is T_0 and the ambient temperature inside the house remains at T_s. An approximation of how long it will take for the wall to cool off can be calculated by assuming T_s does not change much overnight. In many applications the parameter λ is experimentally determined, however we can obtain an approximate value from the definitions in equation (2.16) where $\lambda = \frac{UA}{mC}$ and U is the U-factor for concrete with mass m and heat capacity C. An approximate value for U for concrete can be found in standard tables, for example from the R-factor value converted to suitable units. For concrete the U value is 7.20 W $(\mathrm{Km}^2)^{-1}$ and $\lambda = \frac{UA}{mC} = 1.5 \times 10^{-5}\ \mathrm{s}^{-1}$ using the size, mass and heat capacity of our wall. If the room temperature remains at 20 °C and our wall heated by 7.5 °C during the day, the time taken to cool back down to room temperature can be calculated from Newton's law of cooling. In this case we have $20 = 27.5\ e^{-\lambda t}$ and the wall provides heat for 5.9 h.

Negative thermal energy storage can also be useful and in fact is one of the oldest forms of thermal energy storage. Before modern refrigeration, snow and ice were often stored in insulated buildings during the winter for use in summer months[6]. A modern application of this is to use surplus electrical capacity to make ice, cold water or chilled earth in an underground area at night during off peak hours to be used for cooling during the day. As noted in table 7.5, the latent heat of fusion of water is 334 kJ kg^{-1} which means that a kilogram of ice can absorb 334 kJ of heat by melting. This method of air conditioning has been used in several places, for example at Stanford University and the Alabama Power Company headquarters in Birmingham, Alabama.

---

[5] The heat capacity equation $C = Q/m\Delta T$ can be written in differential form for incremental heat exchange as $dQ/dt = mC dT/dt$ and combined with Newton's heat transfer equation (2.16), $dQ/dt = UA\Delta T$, to form the differential equation $-\frac{UA(T(t) - T_S)}{mC} = \frac{dT(t)}{dt}$ which has the solution $T(t) = T_S + (T_0 - T_S)e^{-(\lambda t)}$.

[6] Ice from Boston in the US was shipped to Cuba for cold drinks starting in the early 1800s.

7.5 Batteries

Each individual cell in a battery comprises two separate chemical reactions. The *anode* becomes negatively charged as electrons accumulate there, while the *cathode* becomes positively charged, where electrons can be absorbed due to a shortage of electrons. The anode and cathode are in chemical contact with each other by means of an electrolyte, which may be a liquid or a solid permeated with a liquid or a gel.

A lead acid battery has a solid lead oxide cathode, a lead metal anode and sulfuric acid (H_2SO_4 and water) as an electrolyte. At the cathode a reduction reaction occurs with an activation energy of 1.685 V that absorbs two electrons,

$$PbO_2 + SO_4^{2-} + 4H^+ + 2e^- \leftrightarrow PbSO_4 + 2H_2O, \tag{7.1}$$

and at the anode an oxidation reaction occurs, giving off the two electrons,

$$Pb + SO_4^{2-} \leftrightarrow PbSO_4 + 2e^-, \tag{7.2}$$

with an activation energy of 0.356 V. The two activation energies add up to about 2 V so a single cell of a lead acid battery produces 2 V of electrical potential. Electrons at the anode are at a higher potential energy than electrons at the cathode and so a current (measured in amperes) will flow from the anode to cathode through an external circuit. Maximum efficiency for the rate of reaction occurs if the cathode and anode are constructed with the largest surface area possible. As the reaction proceeds, lead and lead oxide slowly dissolve into the sulfuric acid. Cells are typically combined to make 6 V, 12 V or 24 V batteries, all multiples of the 2 V primary reaction.

The chemical reactions in equations (7.1) and (7.2) are reversible, so if current from some external source is introduced to the battery and resistive supplies electrons to the cathode and removes them from the anode the reaction will run in reverse, thus recharging the battery. During this process lead and lead oxide deposit back onto the anode and cathode. A factor that limits the cycle life of a lead battery is that the deposition occurs randomly so that the physical configuration of the anode and cathode do not maintain the same surface area as when they started. Some lead also precipitates out of the solution to the bottom of the cell rather than back onto the anode.

The chemical reactions shown above also give off heat while discharging and absorb heat while charging. This heat exchange due to entropy changes is usually overwhelmed by internal resistance heating given by equation (2.18). For the lead acid battery the resistive heat flow is relatively small but for the reactions in some types of lithium-ion (and other) batteries the resistive heat flow is significant if the discharge occurs fast enough. This has been a problem for some cell phone batteries, for example the Samsung Galaxy Note 7 released in 2016 which was recalled by the manufacturer. In the event that the battery is damaged, the heat released from uncontrolled reactions can result in a fire as occurred for a few Tesla electric cars in 2013[7]. Some battery reactions, for example the lead acid reactions, also tend to give

---

[7] The Toyota Prius and some other hybrid and electric cars use NiMH batteries instead of lithium-ion to avoid the potential fire hazard, although the energy density is one third that of the lithium-ion battery.

Table 7.6. A few common battery types. NiCad is nickel cadmium, NiMH is nickel metal hydride, $ZMnO_2$ is the common alkaline battery, $LiCoO_2$ is the typical lithium battery and $LiMn_2O_4$ is lithium, manganese oxide.

Type	Year commercialized	Energy density (kJ kg^{-1})	Power density (W kg^{-1})	Cell voltage (V)	Cost (\$ $(kWh)^{-1}$)
Lead acid	1881	140	180	2	50–150
NiCad	1956	140–220	150	1.2	400–800
NiMH	1990	100–300	250–1000	1.2	250
Zinc air	1997	720	80–140	1.15	80
$ZMnO_2$	1949	300–700	50	1.5	2000
$LiCoO_2$	1992	320–500	740	3.7	300
$LiMn_2O_4$	1999	570	1800	3.6	300

off hydrogen gas in the recharging part of the cycle which reduces efficiency and is potentially dangerous. These energy losses imply the efficiency cannot be 100% for a discharge/recharge cycle for most batteries.

The reactions shown for the lead acid battery are typical but are not the only possible chemical reactions which produce electrons, as shown in table 7.6. A grade school science fair project that was popular in the 1960s was to make a battery using a zinc strip as the anode and a copper strip as the cathode and a potato or a lemon as the electrolyte. Hundreds of other chemical combinations have been found to produce useful current flows, including some plastics and several organic compounds. Other battery chemical combinations which have been investigated include NaNiCl, Li–Ph, $LiNiCoO_2$, $LiNiO_2$ and Li_2S_8.

After cost, the two determining factors for battery choice in a particular application is energy density in kJ kg^{-1} and power density in W kg^{-1}. As can be seen from the properties listed in tables 7.5 and 7.6, different chemical reactions have different power ratings as well as cycle efficiencies and lifetimes. Lithium-ion batteries are three times more expensive than lead batteries but have three times the energy density and four times the power density, making them favorable choices for both electric cars and personal mobile devices.

As an example of energy storage consider the global average annual electrical energy consumption per household which is about 3100 kWh (13 000 kWh in the US). This is 11 GJ of electrical energy use per household (47 GJ in the US). Storing a one day supply of electrical energy would require storing 30 MJ on average (130 MJ in the US) or 214 kg of lead acid batteries (930 kg in the US). The average car battery weighs about 17.7 kg, so a day's worth of energy could be supplied by 12 batteries per person on average globally, or 54 batteries in the US. Using lithium batteries would cut the number of batteries by a factor of three. An instructive exercise is to calculate the quantity of batteries needed to store several days' worth of energy for a typical city.

Commercial batteries are also sometimes rated in terms of ampere hours, which is the amount of charge stored in the battery and gives an indication of how long the battery will last at a given current output. A 100 Ah, 12 V lead acid battery of the type used in automobiles contains $100\text{ Ah} \times 3600\text{ s h}^{-1} = 3.6 \times 10^5\text{ C}$ of charge.

A volt is a joule per coulomb so at 12 V this represents 4.3×10^7 J of stored energy. If this is being used at the rate of 1.0 kW we have 4.3×10^7 J/1.0 kW = 43 000 s = 11.9 h.

Most electrical applications use alternating current whereas batteries produce direct current so for many applications a conversion from direct to alternating current is required, which adds additional equipment and cost to a battery storage system. Conversion processes are never 100% efficient so battery storage systems have to be designed for capacities larger than the amount required by the application for which they are used. On the other hand, batteries work well in combination with PV solar cells which produce direct current (DC).

7.6 Flow batteries

Flow batteries work very much like fuel cells (discussed in chapter 3) except the chemical reactions are reversible. In most applications two different reactive liquid electrolytes are passed from separate tanks past a proton exchange membrane where they react. Protons pass through the membrane to complete an oxidation–reduction reaction supplying electrons to an anode in one of the electrolytes and absorbing electrons at a cathode in the other electrolyte. A typical reaction that has been investigated is the vanadium redox battery, which uses different chemical states of vanadium on both sides of the cell. On the positive side of the cell VO^{2+} ions are converted to VO_2^+ ions by the exchange of protons through the membrane. On the negative side of the battery V^{3+} ions are converted to V^{2+} ions. Vanadium compounds are dissolved in sulfuric acid to form the ions on each side. The combined activation energy is 1.4 V and the energy density per kilogram of electrolyte is 90 kJ kg^{-1}, comparable to compressed air (below). Adding electrons via an external circuit reverses the reactions and proton flow across the membrane, returning the electrolytes to their original state. Cycle efficiencies for this reaction are on the order of 65% to 75% and the reaction occurs fast enough to be useful for electrical load leveling in a power grid.

The power of the flow battery depends on the size and configuration of the electrodes but the net amount of electricity produced (or stored while charging) is dependent on the volume of electrolytes. This means the amount of stored energy is limited only by the size of the containment tanks which can be increased relatively easily. Having the electrolytes separated also means there is less likelihood of an accidental chemical reaction occurring. The arrangement is less attractive as a mobile fuel because of the low energy density but, on the other hand, the chemicals can quickly be replaced making recharging fast, unlike conventional batteries. Because the process can be scaled up to be as large as desired it is appealing for large scale storage of renewable energy for use in the electrical grid. There are many other chemical reactions being investigated, including using gas instead of liquid for one of the electrolytes and organic compounds which have been developed in the last few years[8]. This field is likely to see rapid development in the next few years as

---

[8] For example quinones found in the so called rhubarb battery (Harvard University 2014).

power companies struggle to integrate intermittent sources into the existing grid (Savage 2015).

7.7 Flywheels

For rotating objects the usual definition of kinetic energy, $\mathrm{KE} = \frac{1}{2}mv^2$, where v is the velocity of the object in meters per second and m is the mass in kilograms, changes to $\mathrm{KE} = \frac{1}{2}I\omega^2$, where I is the moment of inertia in kg m^2 and ω is the angular velocity in radians per second. The moment of inertia measures how much mass an object has and the distance of this mass from the axis of rotation which can be calculated from $I = \int_0^R r^2\,\mathrm{d}m$, where R is the radius of the object and $\mathrm{d}m$ is an incremental piece of the object's mass. Table 7.7 lists the calculated moment of inertia for several shapes and different axis.

As an example, consider a solid disk of radius 1 m with a mass of 2000 kg spinning at a 1050.0 rad s^{-1} (10 000 rpm) for 15 min. The energy stored is $\frac{1}{2}I\omega^2 = \frac{1}{2}(\frac{1}{2}2000\ \mathrm{kg}(1\mathrm{m})^2)(1050\ \mathrm{rad\ s}^{-1})^2 = 551$ MJ. The power available from the device over a 15 min period is 0.6 MW. A collection of several such devices is sufficient for compensating power fluctuations in an electric power grid while devices with longer startup times, such as gas turbines and diesel generators come online. In load leveling applications the flywheel stores energy when electricity is fed into a motor attached to the flywheel and energy is retrieved by using the motor as a generator as the flywheel spins down.

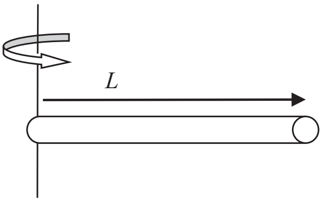
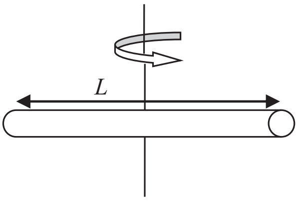
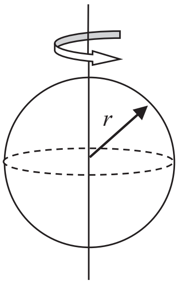
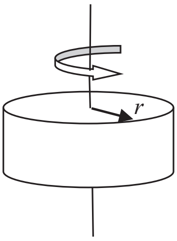
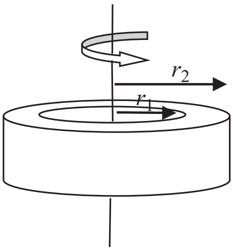
Notice that doubling the radius or angular velocity increases the energy stored by a factor of four. The strength of the material from which the device is made limits rotational speeds to 50 000 rpm for carbon fiber, the strongest material currently available. Modern flywheel applications use frictionless magnetic bearings and mount the flywheel in a vacuum to reduce drag. Shielding for the possibility of disintegration of the rotor increases the weight of large flywheels, limiting their use to stationary applications, although a flywheel propelled bus was built as early as 1940 in Switzerland. The bus had a range of about 10 km (less when full of commuters) at 30 km h^{-1} and was charged periodically when it stopped to pick up passengers. There have been other experimental flywheel-driven vehicles but currently none are in commercial use.

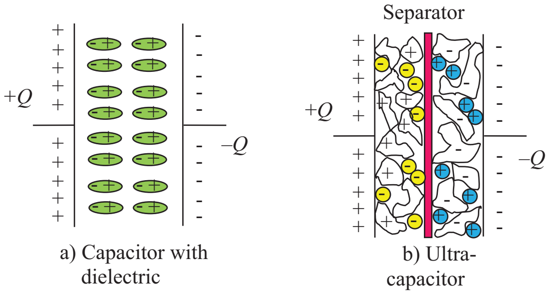
7.8 Capacitors

If two surfaces are place close to each other but not in contact, positive charge on one will maintain negative charge on the other surface allowing for the storage of charge. This arrangement is called a *capacitor*. When the two surfaces are connected through an external circuit, current will flow until the charge is neutralized. Capacitance, in farads, is defined to be the ratio of the amount of charge stored in coulombs to the voltage applied or

$$C = Q/V. \tag{7.3}$$

Table 7.7. Moments of inertia for several regular solids. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

Description and axis	Figure	Moment of inertia
Rod of length *L*(around end)		$I = mL^2/3$
Rod of length *L*(around middle)		$I = mL^2/12$
Solid sphere		$I = 2mr^2/5$
Solid disk (around axis)		$I = mr^2/2$
Cylindrical shell (around axis)		$I = m(r_1^2 + r_2^2)/2$



a) Capacitor with dielectric
b) Ultra-capacitor

Figure 7.4. A cross section of a standard parallel plate dielectric capacitor (a) and an ultra-capacitor (b). From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

The capacitance for any particular physical arrangement is fixed. For a system of two flat surfaces of plates of area A separated by a distance d the capacitance is

$$C = \frac{\epsilon A}{d}, \tag{7.4}$$

where ϵ is a constant called the *permittivity*, which depends on the material between the two surfaces. For parallel plate capacitors the separation between the plates is constant and the electric field, defined as the force on a small test charge, $\vec{E} = \lim_{q\to 0}\frac{\vec{F}}{q}$, will be uniform between the two surfaces. Parallel plate capacitance can be increased by increasing the surface area, decreasing the distance between the surfaces or changing the material between the surfaces. The lowest permittivity is that of a vacuum, $\epsilon_0 = 8.85 \times 10^{-12}\ \mathrm{F\ m^{-1}}$. Dielectric materials which react to the electric field between the plates by becoming polarized have larger permittivities. Figure 7.4(a) is a schematic of the arrangement of molecules between the plates of a charged capacitor. The orientation stores energy by stretching and aligning the molecules of the dielectric. As the charge is removed from the capacitor the molecules reorient, giving up stored potential energy which is available for higher voltage current flow.

The effective surface area of a capacitor can be increased by using a dielectric made of porous material. This arrangement, sometimes called a super- or ultra-capacitor, can hold a thousand times more energy than a regular capacitor of the same size. A schematic is shown in figure 7.4(b).

As stated above, electric fields, E, measured in N/C provide a force on charged particles: $q\vec{E} = \vec{F}$. The work done in moving a charge in an electric field is then $W = -\int_{x_\mathrm{i}}^{x_\mathrm{f}} \vec{F}\cdot \mathrm{d}\vec{s} = -q\int_{x_\mathrm{i}}^{x_\mathrm{f}} \vec{E}\cdot \mathrm{d}\vec{s} = -qE\mathrm{d}$ if the field is uniform. Analogous to gravitational potential energy, the energy stored from moving the charge against the field is $U = qV$, where V is the electrical potential in volts. The work needed to charge (and therefore the energy stored in) a capacitor to a total charge Q is

$$W = U = \int_0^Q V\mathrm{d}q = \int_0^Q \frac{Q}{C}\mathrm{d}q = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}CV^2 = \frac{1}{2}VQ. \tag{7.5}$$

As shown in table 7.5, super-capacitors have very high cycle lives and cycle efficiencies. They also have high power density, but even very large capacitors cannot store very much energy. They are often used for load leveling in electrical grids on very short time scales and have been considered for energy storage during regenerative braking in automobiles.

7.9 Compressed air

Compressed air can be used to drive a piston and so constitutes an energy storage possibility. Compressed air energy storage systems (CAES) are used in a few locations around the world and car manufacturers in Europe and India have built prototypes of automobiles using compressed air. Compressing air heats it and expanding air cools so the process will exchange energy with the environment and thus not be 100% efficient.

The energy which can be stored in a tank of air can be calculated from the work done in compressing a piston: $W = \int_{V_\mathrm{i}}^{V_\mathrm{f}} P\mathrm{d}V$. Using the ideal gas law, $PV = nRT$, where n is the number of moles of gas, R is the ideal gas constant and T is the temperature, we can change this equation to

$$W = nRT \int_{V_\mathrm{i}}^{V_\mathrm{f}} \frac{\mathrm{d}V}{V} = nRT \ \ln\left(\frac{V_\mathrm{f}}{V_\mathrm{i}}\right), \tag{7.6}$$

where we have assumed the process was isothermal (the temperature did not change). The process is not adiabatic, which means a significant amount of heat is lost but this loss could be recaptured using a thermal bath at the cost of making the device significantly heavier. Since we really want the work done in changing the pressure in a tank of constant volume, rather than the volume change at constant pressure, we can use the ideal gas law again at constant temperature to write $P_\mathrm{i}V_\mathrm{i} = nRT = P_\mathrm{f}V_\mathrm{f}$, which gives $\frac{P_\mathrm{i}}{P_\mathrm{f}} = \frac{V_\mathrm{f}}{V_\mathrm{i}}$. For this case the work becomes

$$W = nRT \ \ln\left(\frac{P_\mathrm{i}}{P_\mathrm{f}}\right) = P_\mathrm{f}V \ \ln\left(\frac{P_\mathrm{i}}{P_\mathrm{f}}\right), \tag{7.7}$$

which is the work done in filling a tank of constant volume V from some initial pressure, P_i (for example atmospheric pressure) to a new, higher pressure, P_f. For a 75 l(0.075 m^3) tank with a compression ratio of 1000 to 1 this represents a stored energy of $W = P_\mathrm{f}V \ln(\frac{P_\mathrm{i}}{P_\mathrm{f}}) = 1000\ \mathrm{atm} \times 1.0 \times 10^5\ \mathrm{Pa/atm} \times 0.075\ \mathrm{m}^3 \times \ln(\frac{1}{1000}) = -5.2 \times 10^8\ \mathrm{J}$ (the work done on the gas is a negative number). For a 50% efficient process (to account for thermal loss) this would be 260 MJ. This is approximately the energy found in 7.5 liters of gasoline.

The calculation in the previous paragraph indicates that a compressed air vehicle is feasible and in fact several companies are investigating the possibility. Tata Motors of India has probably done the most research on compressed air cars and has developed prototypes, but has yet to bring a vehicle to market. The earliest large scale CAES system was a 290 MW storage system in Huntorf, Germany, built in

1978. Several such systems have been built or are under construction in Europe and the US.

7.10 Superconducting magnets

Normal wires lose power in the form of heat in proportion to the current squared, $P = RI^2$, as discussed in chapter 2 (equation (2.18)). The motivation for using alternating current is that transformers (which only work with alternating current) can step voltage up and current down, decreasing loss over long range transmission lines. In 1911 it was discovered that some materials will allow resistance free current flow at very low temperatures (below 4.2 K). Superconductors that function above 90 K were discovered in 1986, making the possibility of zero resistance wires chilled by relatively inexpensive liquid nitrogen which boils at 77 K.

Electromagnets made with superconducting wires hold the possibility of storing energy in magnetic fields with very little power loss. The cycle efficiencies of experimental superconducting magnetic energy storage (SMES) are as high as 85% with basically an infinite cycle life. Several SMES systems have been built and used as load leveling devices in the electric grid and the technology is being further developed. Energy per cubic meter stored in a magnetic field is

$$u = B^2/2\mu_0, \tag{7.8}$$

where B is the magnetic field in tesla and $\mu_0 = 4\pi \times 10^{-7}$ Tm A^{-1} is the permeability constant. A 10 T field thus stores $u = B^2/2\mu_0 = 4.0 \times 10^7$ Jm^{-3}.

Superconductors are classified as type I, which lose their capacity for resistanceless current flow at fields of around 0.05 T, and type II, which can maintain magnetic fields up to 40 T. These limits on magnetic field strength restrict the energy that can be stored in an SMES system. A second limitation is the energy investment in keeping these magnets below their critical temperatures. The quick response, high cycle efficiency and cycle life, however, make SMES very useful for short term energy storage.

Projects

1. For each of the energy storage systems discussed in this chapter (flywheel, compressed air, capacitor, etc), calculate the mass and volume of a system of the necessary size to replace a gasoline engine with a 45 l tank. Based on your calculations, which systems have potential for transportation energy storage?
2. Write a report detailing the evolution of the history and current status of all-electric vehicles. Be sure to mention the EV1 produced by General Motors and why that project was abandoned.
3. Calculate the storage necessary to supply all the electricity for your house, city and country for a week. How many lead storage batteries would be required?
4. Based on the data in figure 7.2 (or similar data for a region or city), estimate how much energy would have to be stored to smooth out the seasonal variations of energy demand. Pick one of the energy storage systems

described in this chapter and estimate how large a system would need to be built to smooth out these seasonal variations of energy demand.

5. Make some reasonable assumptions about Trombe walls or other solar thermal absorption systems and design a home or other building for the region where you live which would be able to provide sufficient heating using stored solar thermal energy. Make appropriate assumptions of heat loss for the region.
6. Write a report on the current state of flow battery technology, including chemical reactions under investigation.
7. Write a report on the current status of designs for a superconducting power grid and energy storage.
8. Design a cooling system for a building that uses electricity to make ice at night which will be used to cool the building during the day and compare this with using the electricity directly to cool the building during the day. Make reasonable assumptions about the coefficient of performance (COP) for the refrigeration process. How much of a price differential between peak and off-peak electricity would be needed to make this plan desirable?

References

Ashby M F 2013 *Materials and the Environment: Eco-Informed Material Choice* (Oxford: Butterworth-Heinemann)

The Economist 2007 The truth about recycling *The Economist* Technology Quarterly, Q2 June 9

EIA (US Energy Information Administration) 2016 *Annual Energy Outlook* www.eia.gov/electricity/data/browser/ (Accessed: 22 September 2016)

Electropaedia 2005 Battery and energy technology http://www.mpoweruk.com/index.htm (Accessed: 22 September 2016)

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Harvard University 2014 Organic flow batteries promise breakthrough for renewable energy http://www.seas.harvard.edu/news/2014/01/organic-mega-flow-battery-promises-breakthrough-for-renewable-energy

Hummel R E 2006 *Understanding Materials Science* (New York: Springer)

Savage N 2015 Liquid assets *Nature* **526** S98

Smil V 1999 *Energies, An Illustrated Guide to the Biosphere and Civilization* (Cambridge, MA: MIT Press)

Smil V 2005 *Energy at the Crossroads* (Cambridge, MA: MIT Press)

Sorensen B 2000 *Renewable Energy: Its Physics, Engineering, Environmental Impacts, Economics and Planning* (New York: Academic)

Tester J W, Drake E M, Driscoll M J, Golay M W and Peters W A 2005 *Sustainable Energy: Choosing Among Options* (Cambridge, MA: MIT Press)

Physics and the Environment

Kyle Forinash III

Chapter 8

Transportation

Energy used for transportation in the US, including transportation of goods, has increased from 18.6% of primary energy used in 1973 to nearly 28% in 2014. This should be compared to 24% of the primary energy used for transportation in OECD[1] countries and 15% in non-OECD countries. Privately owned passenger vehicles consume more than 60% of the transportation energy in the US, about 50% in OECD Europe, and around 25% in China and India (EIA 2016). More than 95% of all transport energy comes from petroleum distilates. Probably the easiest and largest energy savings in the US would be to improve transportation efficiency (Richter 2010). The details of our current transportation system are discussed in the following sections from an energy perspective.

8.1 Energy needed to accelerate and move a vehicle

For common shapes traveling at normal transportation speeds in air, the aerodynamic drag force is given by

$$F = \rho A c_\mathrm{d} v^2/2 \tag{8.1}$$

for a vehicle traveling at speed v, where A is the frontal area of the vehicle and ρ is the density of air. Other equations apply to objects moving at very high speed or in more viscous fluids. The drag coefficient, c_d, is a function of the exact shape and roughness of the surface of the object and has no units. Drag coefficients can be calculated for a few simple shapes but the values for most objects are determined empirically in wind tunnel tests. Typical values for vehicles range from 0.35 for most modern cars to 0.15 for laboratory prototypes. We can find the power needed to overcome this drag force at a constant speed by multiplying by velocity, since there

---

[1] The Organization for Cooperation and Development, which includes 35 developed countries, mostly in North America and Europe, plus Japan, South Korea, New Zealand and Australia.

doi:10.1088/978-1-6817-4493-3ch8

is an expression of power given by $P = Fv$.[2] The power to overcome this drag force is then

$$P_{\mathrm{drag}} = \rho A c_{\mathrm{d}} v^3/2. \tag{8.2}$$

The reader can verify that for a typical car with a drag coefficient of 0.3 traveling at 20 m s^{-1} (about 45 mph) and a frontal area of 2 m^2, the power needed to overcome the drag forces will be about 3000 W, which is about four horsepower. This is obliviously a small fraction of a typical car engine's capacity.

Rolling resistance is given by

$$P_{\mathrm{rolling}} = c_{\mathrm{r}} mgv, \tag{8.3}$$

where m is the mass of the car, g is the acceleration of gravity and c_{r} is the coefficient of rolling friction, which is typically around 0.01. For a 1500 kg car traveling at 20 m s^{-1} this is also about 3000 W. Rolling resistance plays a role at low speeds but because drag forces are proportional to velocity cubed, drag is a much bigger force as speed increases. Increasing the speed by 10 m s^{-1} using the previous numbers increases the power needed to overcome rolling friction to 4400 W, but increases power needed to overcome drag to almost 10 000 W. For this reason driving even a little slower greatly improves gas mileage.

The angle of an incline, θ (referred to as the grade), determines the power needed to climb a hill. The engine is overcoming gravitational potential energy, $U = mgh$, where h is the height of the hill. At constant speed the power needed is

$$P_{\mathrm{hill}} = mgv \sin\theta, \tag{8.4}$$

where we have used the fact that $h/t = v \sin\theta$ for a car climbing a hill of angle θ with speed v. To propel a 1500 kg car up a hill of 5° at a constant 20 m s^{-1} requires 26 000 W (35 hp).

For a vehicle which is accelerating the power needed is

$$P_{\mathrm{acc}} = Fv = mav, \tag{8.5}$$

where a is the acceleration. Here if v is the instantaneous speed P_{acc} is the instantaneous power, or if v is the average speed in an interval P_{acc} is the average power in that interval. To accelerate a 1500 kg vehicle from zero to 100 km h^{-1} (28 m s^{-1}) in 5 s (acceleration of 5.6 m s^{-2}) requires an average power of around 120 000 W (about 160 hp). Achieving this speed at an acceleration of 1 m s^{-2} would require 28 s and a power of 21 000 W. Clearly high powered engines are only needed if one wishes to accelerate rapidly. From this equation we can also see that more power is needed to accelerate from a high speed than from a lower speed. To achieve an acceleration of 1 m s^{-2} while moving at an average speed of 10 m s^{-1} requires a power of about 15 000 W for a 1500 kg car, but the same acceleration of 1 m s^{-2} with an average speed of 50 m s^{-1} would require a power of 750 000 W.

---

[2] Power is $P = \mathrm{d}W/\mathrm{d}t = \int_{x_{\mathrm{i}}}^{x_{\mathrm{f}}} \vec{F} \cdot \mathrm{d}\vec{s}/\mathrm{d}t$, so for a constant force and with $\mathrm{d}\vec{s}/\mathrm{d}t = \vec{v}$ we have $P = \vec{F} \cdot \vec{v}$.

Gasoline motors have low efficiencies, typically less than 25%, due to the second law of thermodynamics, as shown in chapter 2. Can cars be made more efficient? Figure 8.1 shows the US mandated Corporate Average Fuel Economy (CAFE) for passenger vehicles and light trucks (which includes pickup trucks and SUVs) and the actual fleet average gas mileage for these vehicles. From 1985 to 2005 for trucks and to 2010 for passenger cars the mandated CAFE standard did not change and the fleet average fuel economy also stayed constant. Instead, new technology was used to make vehicles larger and faster, increasing average vehicle size by 9% and average horsepower by 40% (the ratio of hp to weight is shown by the black dotted curve). The mandated standard was raised in 2010 for passenger cars and is scheduled to reach more than 50 mpg (21 km l^{-1}) by 2025. Light truck fleet average is mandated to be 40 mpg (17 km l^{-1}) by 2025.

How can cars become more fuel efficient? The mass of the car has a linear effect on the power needed to accelerate, go uphill and overcome rolling resistance, so reducing the mass of a vehicle will save energy by reducing the power needed. Turning the engine off while idling also has an effect on gas mileage and many cars already have this feature. Power needed to overcome drag is proportional to speed cubed and this is a more significant effect than weight. In the mid-1970s the US lowered highway speed limits from 70 mph (113 km h^{-1}) to 55 mph (88 km h^{-1}) due to a global energy crisis. This reduced the power needed to maintain a steady speed by 50%, greatly improving gas mileage, which is linearly proportional to power used. From these calculations we see that gas mileage can be improved by accelerating more slowly, coasting, anticipating the traffic ahead to reduce the amount of braking and thus the amount of acceleration needed and driving at slower speeds. These techniques (sometimes called hypermiling) can result in a doubling of the rated gas mileage of a normal car. European countries impose a high gasoline tax (nearly doubling the price of a liter of fuel compared to the US) to motivate

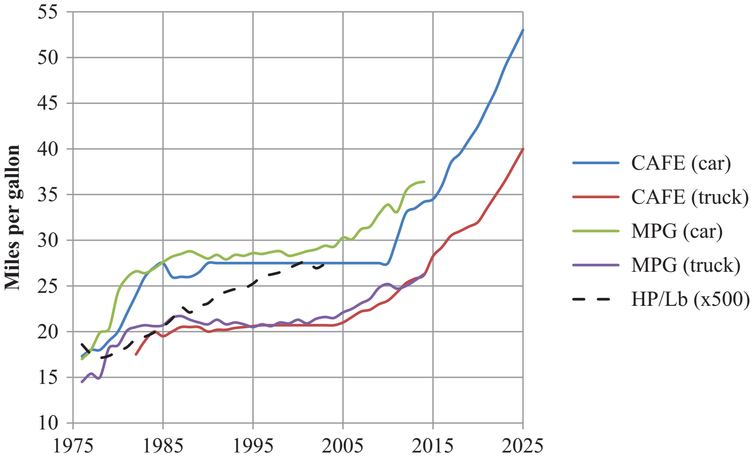


Figure 8.1. CAFE standard and actual fleet average gas mileage of vehicles in the US. Trucks are light, personally owned vehicles such as pickup trucks and include SUVs. (The fate of these mandates is in question under the present US administration at the time of this writing.)

drivers to save fuel by driving smaller cars and practicing fuel saving driving behavior and this is one route to using less transportation fuel.

Electric motors are more efficient than gasoline engines, especially at low rotation speed where they have high torque. Because of the low energy density of batteries which limit the range of all-electric cars and the long times needed to recharge batteries, these vehicles and have so far not been popular. A switch to electric vehicles would reduce the total primary energy used for transportation because electric motors are more efficient. The net efficiency of a 90% efficient electric motor and a 60% efficient gas turbine electric generating plant uses less primary energy than a 25% efficient gasoline engine.

Hybrid cars combine the advantages of electric motors with the longer range of gasoline engines. The gasoline motor can either act to only recharge a battery which runs the electric motor or run in tandem with the electric motor with a computer system to switch back and forth as appropriate to the driving conditions. The larger batteries in hybrid and all-electric vehicles can also recapture some of the kinetic energy when the car is brought to a stop. Ordinary cars waste the forward kinetic energy of motion when they stop by turning this energy into heat in the break-pads. If instead this energy is used to drive a generator to recharge a battery it can be used to power the acceleration of the vehicle via an electric motor and the overall process is more efficient. A 1500 kg car traveling at 20 m s^{-1} has a kinetic energy of 0.3 MJ. If half of this energy can be captured by making 100 stops from 20 m s^{-1} the energy saved is 15 MJ, equivalent to the energy contained in half a liter of gasoline.

8.2 The energy cost of energy

When a primary energy source, such as petroleum, is converted into a usable fuel, some energy becomes unusable in the conversion process due to the second law of thermodynamics. Table 8.1 lists several conversion processes of a primary source

Table 8.1. Conversion efficiencies for various end fuels (Smil 1999). These are different types of conversions, for example chemical to chemical (petroleum to gasoline), electromagnetic waves to electricity (PV) and physical (natural gas to compressed natural gas).

Process	Conversion efficiency, %
Natural gas to compressed	85
Crude oil to gasoline	85
Natural gas to electricity (gas turbine)	65
Natural gas to hydrogen	60
Coal to gasoline (Fisher–Tropsche)	50
Coal to electricity (conventional)	35
Grid electric to hydrogen	24
Solar to electric (PV)	15–25
Soybeans to bio-diesel	40
Corn to ethanol	20
Plant photosynthesis to bio-mass	<8

into a usable fuel. These losses are only a small part of the calculation of EROEI (see chapter 4); energy is also used in the extraction process itself as well as in the transportation of the fuel to the market. From the table we see that approximately 15% of the energy of crude oil is lost in converting it to gasoline for use in an automobile and about the same amount is lost in compressing natural gas into tanks for commercial use.

We can expect that some of these efficiencies, for example PV solar cells, will improve over time with advances in technology. However some processes in table 8.1, for example making hydrogen from natural gas or electricity, are probably at or close to the best conversion efficiency possible.

Calculating the efficiency of a multistep process requires multiplying the efficiencies of each step so that the lowest efficiency step dictates the overall efficiency of the process, as we saw in chapter 2. A calculation of EROEI for producing gasoline, for example, would multiply the conversion efficiency of 85% from table 8.1 by the extraction efficiency, which is about 95% for a new conventional oil well. The extraction efficiency decreases over time, however, as more energy intensive extraction methods such as steam injection and hydraulic fracturing are applied. Similarly, soybeans are about 1.5% efficient in converting sunlight to plant material so the overall efficiency of converting sunlight to soybeans to biodiesel is $0.40 \times 0.015 = 0.006 = 0.6\%$, not including the energy investment in the form of fertilizer and farming equipment needed to grow the soybeans.

The Fisher–Tropsche process, invented in Germany in 1920 and used during the Second World War to supply the German army with fuel, starts with coal as a primary source and produces hydrocarbons which can be used to make a portable fuel. Coal or other biomass is first heated and partially combusted to give carbon monoxide in the following illustrative processes:

$$\mathrm{C} + \mathrm{H_2O} \rightarrow \mathrm{H_2} + \mathrm{CO} \quad \text{(from coal)}, \tag{8.6}$$

or

$$2\mathrm{CH_4} + \mathrm{O_2} \rightarrow 4\mathrm{H_2} + 2\mathrm{CO} \quad \text{(from methane)}. \tag{8.7}$$

The carbon monoxide is then combined with hydrogen to make liquid hydrocarbons,

$$(2n+1)\mathrm{H_2} + n\mathrm{CO} \rightarrow \mathrm{C}_n\mathrm{H}_{(2n+2)} + n\mathrm{H_2O}, \tag{8.8}$$

where n is a whole number and resultant compound is a complex hydrocarbon molecule. These compounds can be refined in much the same way as petroleum to provide a liquid fuel. The overall efficiency of the process is about 50%, as given in table 8.1, because of the input energy needed to make the process occur.

For both biodiesel and fuel from the Fisher–Tropsche process, using the primary energy source (soybeans and coal) directly in an engine is problematic. This, then, is a potential reason for using these processes even though the energy losses are relatively high; they provide a portable fuel. As we will see below, other factors involved are the efficiencies of the engines being used and the pollution produced.

8.3 Using hydrogen for transportation

Hydrogen can be burned in an ICE but the efficiencies are limited by the second law of thermodynamics. Using hydrogen in a fuel cell, however, yields a much higher efficiency than most thermal processes. Hydrogen has a very high energy content per weight making it an attractive fuel and several fuel cell based prototype cars have been built. It is important to realize, however, that hydrogen does not occur in a free form, even though it is the most abundant element in the Universe. On Earth all hydrogen is found bound to other elements in compound form and energy is required for its release.

There are several potential sources for hydrogen, but more than 95% of the commercially available hydrogen is derived from fossil fuels (mostly natural gas) using a process called steam reforming. Steam at high temperatures (around 800 °C to 1000 °C) is used to break down methane into hydrogen and carbon monoxide:

$$\text{energy} + CH_4 + H_2O \rightarrow CO + 3H_2. \tag{8.9}$$

This process is about 60% efficient because of the energy involved in making the steam. Unlike the case for coal or soybeans, here we have converted a portable fuel, methane, into a second portable fuel, hydrogen, with a 40% loss in primary energy. This only makes sense from an energy perspective if the hydrogen can be used in a fuel cell at a much higher efficiency than the methane. The total efficiency of a 90% efficient electric motor run from electricity from a fuel cell with 60% efficiency and a fuel reforming efficiency of 60% would be 32%. Burning methane in an ICE has an efficiency of 15%–25% so in terms of overall efficiency a switch to fuel cells run on hydrogen does make sense but burning it does not. However it is also the case that some types of fuel cells can run directly on methane which might achieve higher overall efficiency if the methane fuel cell efficiency is high enough.

It is also possible to generate hydrogen from electricity, which would be very attractive if the electricity was generated by renewable energy. In electrolysis direct current is applied to water and the electrical energy separates the water into hydrogen and oxygen in gaseous form. Hydrogen is formed at the anode (positive electrode) and oxygen forms at the cathode (negative electrode). The electrodes are made of platinum or some other catalytic metal and the two gases are trapped at each electrode before they can recombine. Currently only the hydrogen is collected for use as a fuel. Because some heat is generated the efficiency of the entire process (output thermal energy obtained by burning the hydrogen divided by input electrical energy) is approximately 60%, not including the energy needed to generate the electricity (higher efficiencies are reported if the energy available in the oxygen is included). Using a gas turbine with 60% efficiency to generate the electricity would result in a hydrogen fuel created at an efficiency of 36%, comparable to the efficiency of making hydrogen from steam reforming. However, using coal generated electricity at 40% efficiency reduces the efficiency to 24%.

The amount of electricity needed for electrolysis decreases as temperature increases because some of the energy needed to break the hydrogen–oxygen bond comes from thermal energy rather than electrical energy. Above 2500 °C water will spontaneously break down into hydrogen and oxygen without any electric current (this process is the source of the hydrogen explosions in the Chernobyl and Fukushima nuclear disasters). Thermal energy is created from fossil fuels with near 100% efficiency so this so called 'high temperature electrolysis' can have very high efficiencies. As was the case for biodiesel and the Fisher–Tropsche process, using fossil fuels to create hydrogen is not more energy efficient but may be desirable if a portable fuel is the objective. Likewise, if the waste heat from fossil fuels or other sources is used for other purposes the combined efficiency can be higher, as in the case of cogeneration. Other potential heat sources for high temperature electrolysis are heat from nuclear reactors, solar thermal energy and geothermal sources.

Another potential source of hydrogen is directly from plants. The sugars and starches made by plants as plant structures such as roots, branches and leaves can be burned as a fuel but enzymes and fermentation processes can also release hydrogen from these plant components. Although plant photosynthesis has a low efficiency with most of the collected solar energy going to plant development, these enzymatic processes can produce a portable fuel. It has long been know that algae and other primitive organisms use photosynthesis as an energy source and will produce methane or hydrogen as a byproduct under certain conditions. Algae are responsible for the origin of the oxygen in the Earth's atmosphere from photosynthesis with water and carbon dioxide as inputs. Some algae can be manipulated to switch to producing hydrogen by reducing the sulfur available to the organism. There are a wide variety of algae types and there has been some success in genetically manipulating them to have higher efficiencies than the usual 1% found in most plants. The chemical process involved has also been investigated and artificial versions of photosynthesis are under investigation as potential hydrogen sources as well as sources of hydrocarbons from CO_2 which could be used as fuel (see Crabtree and Lewis 2007 and Liu *et al* 2016).

Containing hydrogen, either as a liquid or compressed gas, is difficult because of the small size of the molecule, as mentioned in chapter 7. The compression or liquefaction process requires energy and the weight of the tank reduces the energy density significantly, as noted in table 7.3. Exposure of high strength steel to hydrogen also eventually weakens the steel which can lead to failure of pipes under high pressure. These properties make hydrogen harder to manipulate than traditional fuels.

There is a range of safety considerations involved in storing and handling hydrogen as compared to other fuels such as methane or gasoline (Huld 2001). Because of its small molecular size and light weight it diffuses faster than other fuels if there is a leak. The temperature at which it will auto ignite without a spark is also higher than methane or gasoline. On the other hand hydrogen burns without a visible flame making a hydrogen fire difficult to perceive and it can ignite explosively at a wider range of concentrations than other flammable gases. A careful risk assessment will be necessary before hydrogen is widely introduced as an alternative fuel. For many of the above reasons some investigators have questioned the desirability of a hydrogen based transportation system (Shinnar 2003).

8.4 Well-to-wheels or life-cycle efficiency

If we wish to replace or modify any system or device based on energy considerations alone, it is important to examine the energy efficiency of the conversion of primary energy into fuel, the fuel energy used by the device over its lifetime and the energy used to make and then dispose of the device at the end of its life. For a transportation vehicle or other transportation method this is referred to as the *well-to-wheels efficiency*; for other devices or products it is often called the *life-cycle efficiency*. Here we look at the well-to-wheels efficiency for several vehicles and their related emissions.

Figure 8.2 shows a series of calculations performed to compare the efficiencies of various combinations of engines and energy sources. The values shown were generated using a free downloadable spreadsheet tool from Argonne National Laboratory which allows the comparison of well-to-wheels energy use and greenhouse gas emissions for hypothetical combinations of fuel sources and engines (ANL 2016). The resulting figures use existing technology and conventional sources but sometimes in combinations that do not currently exist, for example hybrid diesel cars. In the figures shown, grid electricity is assumed to be generated from a mix

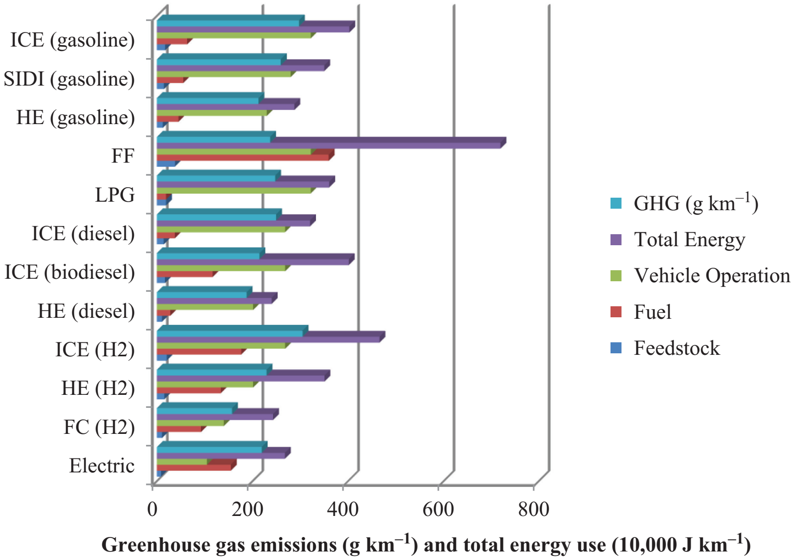


Figure 8.2. Well-to-wheels greenhouse gas (GHG) emissions in grams per kilometer driven and energy use for several choices of alternative vehicles (data from the GREET simulation available from Argonne National Laboratory (ANL 2016)). 'Feedstock' refers to the energy needed to extract and transport the raw fuel, 'Fuel' refers to the energy needed to convert the feedstock to a usable fuel and 'Vehicle operation' refers to the energy actually used in normal vehicle operation during its lifetime. The following vehicles are shown: ICE (internal combustion engine), advanced technology engine (SIDI), hybrid engine (HE), flexible fuel vehicle using 85% ethanol (FF), liquid natural gas engine (LPG), fuel cell (FC), and all-electric. Fuels considered are hydrogen (H2), diesel, gasoline and liquid natural gas (LPG). From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

of 50% coal fired plants, 18% natural gas, 20% nuclear and the rest from other sources such as wind and hydroelectric (these proportions can be adjusted in the spreadsheet). Hydrogen is assumed to come from steam processing of natural gas with efficiencies as described above. The figures assume the energy cost of manufacturing the vehicle, disposing of the vehicle at the end of its useful life and its useful lifetime are approximately the same for each vehicle and so are not included. We can expect new technology to change some of the assumptions of the model, but it can at least give some basic guidance for reasonable choices for transportation in the near future.

The well-to-wheels energy use in figure 8.2 is measured in J km^{-1} and greenhouse gas emissions in g km^{-1}. The figures show that hybrid technology improves gas mileage, total well-to-wheels energy use and greenhouse gas emissions compared to conventional ICEs. Hybrid diesel, hydrogen fuel cells and all-electric vehicles (using grid electricity) have similar total efficiencies and emissions. These vehicles seem to be reasonable choices for further development, at least from an energy and emission perspective. Vehicles built to run on various mixtures of alternative fuels such as biodiesel or ethanol (flex fuel vehicles) are the net losers in this comparison, both in terms of energy used in vehicle operation, total energy used and emissions. There is currently a shift in generating grid electricity in the US towards cleaner and more efficient natural gas because of the low prices as the result of hydraulic fracturing. This is likely to change the results shown in figure 8.2 for electric and plug-in hybrids, bringing further improvements in total fuel economy and emissions control. There are a large number of options which can be explored using the GREET model; the reader is encouraged to download the software and experiment with alternative well-to-wheel combinations.

On the surface it would seem that all-electric vehicles would be a reasonable choice as a transportation mechanism because of the high well-to-wheels efficiency. Several practical concerns should be raised, however. The US uses about 27 EJ of primary energy (mostly petroleum products) annually for transportation in vehicles with less than 25% efficient motors. This means only about 6.7 EJ of this energy is actually used to move the vehicles. For a 70% efficient electric motor with a 70% efficient battery storage system only 13.8 EJ of electrical energy would be needed to supply this 6.7 EJ of mechanical energy. If we assume a typical electric power plant can produce 1000 MW and run 80% of the time, this plant can provide 0.025 EJ of electrical energy in a year's time. To supply the additional 13.8 EJ of electricity needed to convert all vehicles to electric would require around 550 new such power plants. This is half again as many existing plants as currently exist in the US (equivalent to 950 plants, each 1000 MW). A 2 MW wind turbine operating at a 50% capacity can produce 3.2×10^{-5} EJ in a year so, alternatively, 430 000 wind turbines could supply the needed electricity. Clearly these are not small changes in infrastructure.

The energy used in and emissions from plug-in hybrids were not included in figure 8.2, in part because it is highly dependent on how the car is driven. Plug-in hybrids can operate entirely as electric cars if they are only driven for short distances or they may function more as an ordinary hybrid if the car is driven over longer distances.

Vehicle owners are more interested in the fuel they have to purchase for the vehicle than well-to-wheels efficiency, however. Although the energy used by a vehicle per kilometer or mile driven can be easily measured or calculated, comparing fuel economy between gasoline engines and all-electric or plug-in hybrid cars is problematic. Electrical processes are inherently more efficient so the same amount of energy purchased as gasoline and electricity does not buy the same usable amount of kinetic energy for the vehicle. The standard in the US, mandated by the EPA, has been to use a figure for the energy content of gasoline (32.1 MJ l^{-1}) to convert electrical energy used in MJ km^{-1} to MPGe (miles per US gallon equivalent). For example, based on laboratory testing, the EPA measures the Tesla Motors' Model S electric car with the 70 kWh lithium battery pack option to have a range of 390 km which is a fuel economy of 0.65 MJ km^{-1}. The Tesla S then would have an equivalent gas mileage of 116 MPGe (49.4 km l^{-1}). The same process can be used to compare the fuel efficiency of vehicles based on hydrogen fuel cells or other energy sources. The calculation, however, makes the untrue assumption that all energy sources have equivalent efficiencies.

The effect of greenhouse gas emissions on the climate will be discussed in chapter 9, however, as figure 8.2 shows, the choice of fuel and engine have an impact on pollution as well as total energy used. The emissions from transportation have both a global impact as greenhouse gases and more localized, direct effects on human health. The well-to-wheels contamination emitted by all-electric vehicles depends strongly on the source of electricity used to charge the batteries. Detailed analysis of various combinations of fuel sources and motor types shows that all-electric vehicles would be among the worst generators of pollutants if the electricity is generated from coal but among the cleanest and healthiest choices if the electricity comes from renewable energy sources or natural gas (Tessum 2014). Vehicles using corn ethanol as a fuel source produce pollutants comparable to electric vehicles using electricity from the current grid mix and these values are almost double that from pollution from conventional gasoline engines. Making a responsible choice among alternatively powered vehicles requires an analysis of both the well-to-wheels energy consumed and the well-to-wheels pollutants emitted and is highly dependent on regionally varying sources of electricity.

The numbers in figure 8.2 do not include the energy needed to build the vehicle and the energy needed to dispose of the vehicle after its useful life. The energy needed to refine the material to make a car varies depending on the type of car, figures which are also available from the GREET model calculations. Conventional internal combustion vehicles, depending on the vehicle, require on the order of 0.4 GJ to construct, with electric vehicles and fuel cell vehicles using light-weight materials requiring as much as 100 times that because of the specialized materials needed. For an all-electric vehicle requiring 50 GJ to construct the equivalent amount of gasoline energy consumed, using 32.1 MJ l^{-1}, would be 1600 l or 400 gal, roughly the average annual fuel consumption for a US commuter in a conventional vehicle with typical gas mileage. For many accounting purposes the usable lifetime of a car has traditionally been stated as eight years, although the actual lifetime has risen to 11, meaning that as a rough estimate the energy needed to build a car is around 10% of the energy it consumes as fuel. This fraction will be higher for vehicles with newer technology because the energy needed to make the vehicle is greater and the fuel used less than the conventional vehicle.

8.5 Transportation alternatives

As mentioned previously, 28% of the primary energy consumed in the US is used for transportation. Globally 44 EJ of gasoline, 40 EJ of diesel, 13 EJ of jet fuel, 10 EJ of fuel oil, 3 EJ of natural gas, and 1 EJ of electricity and other liquids were used for transportation in 2012, representing 20% of the total primary energy used that year. Energy used for various transportation alternatives in the US are compared in table 8.2. Not shown are walking, which requires about 0.20 MJ in the form of food per kilometer and cycling, which is more efficient at around 0.15 MJ km^{-1}. Many non-self-propelled means of transportation take ten times as much energy per kilometer traveled.

The figures in table 8.2 can be misinterpreted in at least two ways. In the US, statistics for trucks includes small pickup trucks which are used exclusively for

Table 8.2. Energy use figures for transportation in the US in 2013 (Davis *et al* 2015). Personal trucks include SUVs. Demand response includes taxis, limousines, Uber, Lyft and other types of on-demand transportation. 'Other' includes agricultural equipment, construction and mining equipment, railroad maintenance equipment, industrial equipment and off road vehicles, such as snowmobiles, all-terrain vehicles and golf carts.

Transportation method	Number of vehicles (thousands)	Passenger km (millions)	Load factor (persons/vehicle)	Energy intensity (MJ per passenger km or MJ per metric tonne km)	Energy use (EJ)	Energy (%)
Cars	113 676.0	3 586 080	1.55	2.07	7.43	28
Personal trucks	106 018.4	3 039 838	1.72	2.31	7.01	26
Motorcycles	8405.0	37 800	1.16	1.63	0.06	<1
Demand response	68.6	3474	1.4	8.02	0.03	
Buses	792.0	35 690	9.2	2.68	0.22	1
Air, commercial	199.9	927 910	105.2	1.58/15.8	1.69	6
Recreational boats	13 706.6				0.26	
Rail					0.58	2
Passenger	20.2	62 485	26.9	1.62	0.10	
Freight				0.22	0.48	
Heavy trucks	26 483.0	432 028		0.92	6.21	20
Pipeline				0.20	1.24	4
Water				0.15	1.02	3
Other					2.25	8

passenger transportation. In most of the rest of the world small trucks are mostly used for the transport of goods and services. Care should be taken in comparing the numbers in table 8.2 to similar figures in other countries. A second difficulty is the figure for air travel. A modern commercial jet uses most of the energy it consumes in reaching a cruising altitude. Fuel economy (energy intensity) per passenger kilometer, then, is highly dependent on how long the individual trip is. A passenger car can beat the fuel economy of a jet for a short trip of 200 km or so but not long trips of thousands of kilometers.

From the figures in table 8.2 it should be clear that when goods (or people) can be put on a boat or in a pipe the energy used per kilometer is much lower; comparable to the fuel efficiency of cycling or walking. Likewise rail is an energy efficient means for moving both passengers and goods compared to cars and trucks. Travel by taxi (listed under demand-response) is energetically unfavorable because these vehicles often continue to circulate empty while they look for passengers. It is possible that new on-demand services like Lyft and Uber will be much more energy efficient since these services do not need to constantly roam in search of clients. Likewise, self-driving cars hold out the potential for a much more compact private transportation system since jointly owned cars could service a larger number of people more efficiently and not require parking space (The Economist 2016). On a first consideration one would think that buses ought to be energetically efficient but, at least in the US, buses used for public transportation are often not very full. The buses have large capacities to accommodate the heavy traffic during rush hour but otherwise may circulate mostly empty.

8.6 Transportation risk

Another factor involved in selecting between transportation alternatives is risk of injury or death. As we will see in chapter 10, getting into a car is one of the more risky things most people do in their life. In 2014 there were 34 568 transportation related fatalities in the US, equivalent to the number of deaths in the terrorist attacks on 11 September 2001 occurring every month. A breakdown of the percent of transportation deaths for each transportation mode is given in figure 8.3.

The numbers in figure 8.3 make it clear that driving your own vehicle, be it passenger car, light truck, motorcycle, bicycle, recreational boat or small private plane, is hazardous. Pedestrians walking near highways are also at risk of being killed in a traffic accident. On the other hand, very few people die on buses, trains or subways, and there were no fatalities at all due to commercial flights in the US in 2014. Clearly mass transit is far safer than personal transportation modes. How these risks compare to other hazards encountered in modern life will be investigated in chapter 10.

Transportation energy use and safety have unexpected connections. Figure 8.4 shows the vehicle miles traveled (VMT) and traffic fatality rate per year in the US starting in 1920. Fatality rates dropped significantly in the 1940s, in part because young male drivers left to go to war but also because of gasoline rationing which greatly restricted domestic travel. The global energy crisis of the early 1970s resulted

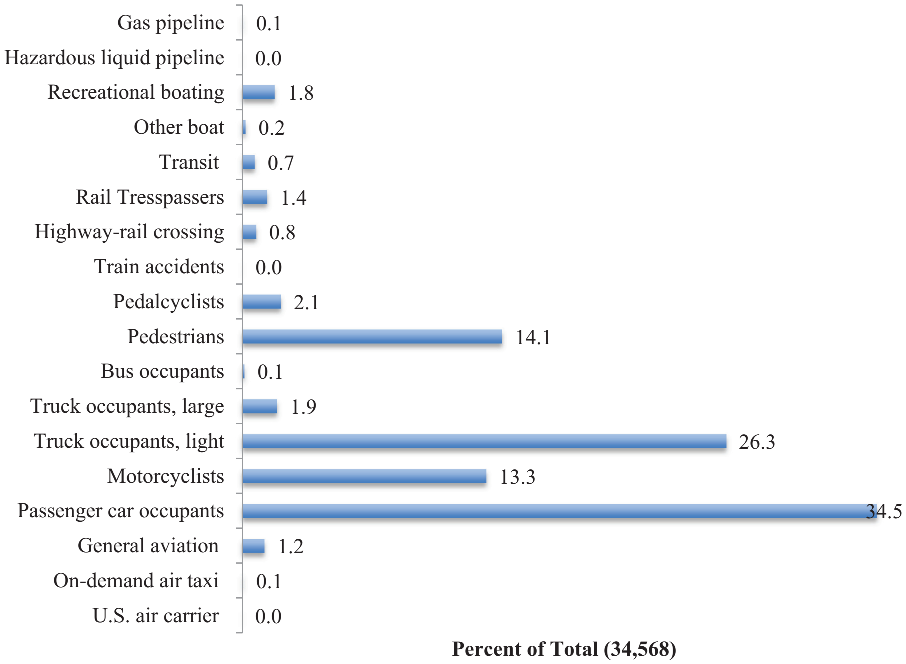


Figure 8.3. Transportation related deaths in the US in 2014 (BTS 2014). 'Other boat' includes commercial boating accidents in US territorial water. Transit includes subways and light passenger rail. General aviation includes small private planes.

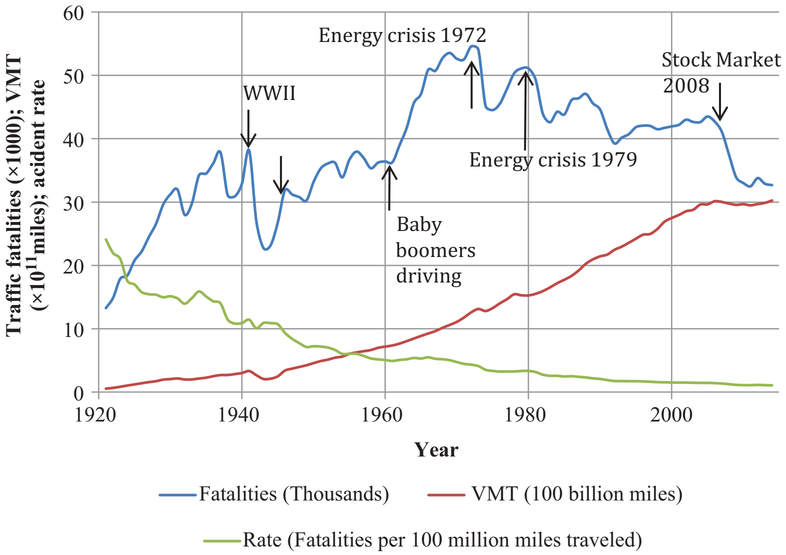


Figure 8.4. US traffic related fatalities in thousands of people, VMT (vehicle miles traveled) times 10^{11} and fatal accident rate per VMT in the US (Wenzel and Ross 2007, BTS 2014). The figures include pedestrians killed by vehicles, and truck and bus accidents.

in a reduction of the speed limit in the US in 1974 to a national limit of 55 mph (88 km h^{-1}), down from as high as 80 mph (128 km h^{-1}) in a few places before the law went into effect. The energy savings were only about half of what was predicted but the unexpected result was a sharp drop in highway fatalities. Oil prices nearly doubled in 1979 due to the revolution in Iran, making gasoline much more expensive resulting in a change in driving habits and a decrease in the fatality rate in the years following. A similar fatality rate decline can be seen as a result of the economic crisis of 2008 which included a leveling off of VMT in the US.

Figure 8.4 also shows that the number of fatalities per 100 million miles traveled has steadily decreased in the US. Similar trends are seen globally under very different driving conditions and in some cases with much poorer roads and older vehicles. Although cars and roadways are becoming safer, the greater impact on the fatality rate is from better driver education and more experience in driving and managing traffic as the number of cars increases. Many less developed countries have much higher fatality rates with a much lower density of cars and drivers, but the fatality rate per kilometer traveled is also seen to decrease as the number of cars, drivers and miles driven increases.

Projects

1. Download the Greet Model from Argonne National Laboratory (https://greet.es.anl.gov/net). First replicate the figures in figure 8.2 and then try various combinations of fuel sources and engine types. Can you find a combination that beats the best efficiency shown in figure 8.2?
2. Write a report on current status of using hydrogen fuel cells for transportation. Include a discussion of existing prototypes.
3. Go back to table 7.3 and pick three or four fuels. For the number of cars in your city (or country or county) calculate how much of each of those fuels would be needed to replace the fuel currently being used assuming the same efficiency as a gasoline engine.
4. Estimate the fuel savings predicted for lowering the speed limit from 80 mph (128 km h^{-1}) to 55 mph (88 km h^{-1}) for cars in the US in 1973. Start with a calculation of the energy needed for the drag forces at these speeds. Make appropriate assumptions about other parameters.
5. Calculate the energy you use during your daily commute. Use estimates of the amount of accelerating, constant speed and hill climbing you do (this can be done for automobiles, buses or subways). Which part(s) of the commute require the most energy? Estimate the energy savings by using a more streamlined vehicle with half the acceleration and regenerative braking where 40% of the energy is captured. Provide calculations for any other energy savings you can think of (for example a modified drag coefficient, car weight etc).
6. Carefully compare your fuel economy during a week driving at the speed limit with a week driving 10 km h^{-1} lower than the speed limit. How much of a saving is this over a year's time?

7. Estimate the primary energy savings if 20% of the cars in the US (or other country) switch from diesel or gasoline to all-electric where the electricity is provided by (a) coal, (b) natural gas and (c) your choice of renewable energy and the engine is 80% efficient.
8. Estimate the primary energy savings if all cars in the US are brought to the 2025 CAFE standard compared to the 2005 standard. Make reasonable estimates about miles driven, number of cars, etc.
9. Analyze the idea of a car powered entirely from solar energy from an energy perspective. Start with current solar panel efficiencies and existing lithium batteries. Make reasonable assumptions of available solar energy and the amount of driving to be done. How much of an improvement in PV and battery efficiency is needed to make this work?
10. Start with references Wenzel and Ross (2007) and Evans (2002) and write a report about how to make driving safer. Are larger vehicles always safer? How much of a role does driver education play in automobile crashes? Make comparisons between accident rates in different countries and speculate on the reasons they are different (include a discussion of driver education requirements in these countries). Make comparisons with the accident rates of driverless cars per mile traveled (if these rates are scaled up to the same number of miles driven by humans, what will the accident rate be?).

References

ANL (Argonne National Laboratory) 2016 GREET Model https://greet.es.anl.gov/net (Accessed: 11 July 2016)

BTS (US Bureau of Transportation Statistics) 2014 table 2-1: Transportation fatalities by mode www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_02_01.html (Accessed: 16 December 2016)

Crabtree G W, Dresselhaus M S and Buchanan M V 2004 The hydrogen economy *Phys. Today* **57** 39

Crabtree G W and Lewis N S 2007 Solar energy conversion *Phys. Today* **60** 37

Davis S C, Diegle S W and Boundy R G 2015 *Transportation Energy Data Book* 34 edn (Oak Ridge, TN: Oak Ridge National Laboratory) ORNL 6991 http://cta.ornl.gov/data/index.shtml

The Economist 2016 Transportation as a service *The Economist* Oct 1, p 58

EIA (US Energy Information Administration) 2016 International Energy Outlook www.eia.gov/forecasts/ieo/pdf/0484(2016).pdf

Evans L 2002 Traffic crashes *Am. Sci.* **90** 244

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

Huld T 2001 Compilation of existing safety data on hydrogen and comparative fuels *European Integrated Hydrogen Project report* www.eihp.org/public/documents/CompilationExistingSafetyData_on_H2_and_ComparativeFuels_S.pdf

Liu C, Colon B C, Ziesack M, Silver P A and Nocera D G 2016 Water splitting–biosynthetic system with CO2 reduction efficiencies exceeding photosynthesis *Science* **352** 2110

Richter B 2010 *Beyond Smoke and Mirrors, Climate Change and Energy in the 21st Century* (Cambridge: Cambridge University Press)

Shinnar R 2003 The hydrogen economy, fuel cells, and electric cars *Technol. Sci.* **25** 455

Smil V 1999 *Energies, An Illustrated Guide to the Biosphere and Civilization* (Cambridge, MA: MIT Press)

Tessum C W, Hill J D and Marshalla J D 2014 Life cycle air quality impacts of conventional and alternative light-duty transportation in the United States *Proc. Natl Acad. Sci* **111** 18490

Wenzel T P and Ross M 2007 Safer vehicles for people and the planet *Lawrence Berkeley National Laboratory publication* LBNL-325E http://repositories.cdlib.org/cgi/viewcontent.cgi?article=6832&context=lbnl (Accessed: 14 December 2016)

Chapter 9

Climate

The Earth's climate is a complex system but one that still must obey the laws of physics. This chapter is based on the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) and explores what is known about the Earth's climate. The IPCC draws together climate experts from around the world to review existing published research on all aspects of climate science including atmospheric science, chemistry, physics and geology. Over 130 countries are represented and thousands of scientists are involved. The material reviewed has already been peer reviewed in order to be published and receives a further evaluation from the specialists on the panels before being synthesized into a formal report. As a result of this rigorous process the IPCC reports represent the most reliable knowledge available at the time the report is published.

9.1 Radiation balance

As discussed in chapter 6 on renewable energy, about 1361 W m^{-2} of energy in the form of electromagnetic waves reaches the top of the Earth's atmosphere. The frequency distribution of this radiation is not uniform but rather is approximately blackbody in shape, as shown in figure 2.3. About 30% of the incoming energy is reflected, 4% by the Earth's surface, 20% by clouds and 6% by the atmosphere. The Earth's surface, then, has an average *albedo*, a, of 0.30 (30%). This value varies depending on the nature of the surface; water absorbs nearly all the incoming radiation while snow reflects most of it. Land surfaces have varying albedos depending on the terrain and this can change, for example if forests are cut and the land is farmed or buildings are constructed.

The surface area of the Earth is $4\pi r^2$ where r is the Earth's radius but, seen from the Sun, only a circular cross section of πr^2 absorbs energy. The total incoming power, P_{in}, from the Sun is then

$$P_{\text{in}} = (1 - a)S\pi r^2, \qquad (9.1)$$

doi:10.1088/978-1-6817-4493-3ch9

where $a = 0.3$ and $S = 1361\ \mathrm{W\ m^{-2}}$. If this energy were to be constantly absorbed the Earth's temperature would continue to rise and, given the age of the Earth, would have resulted in a molten planet. This does not happen because the Earth re-radiates the same amount of energy back into space. The Earth's energy balance varies only slightly from equilibrium over time with roughly the same amount of energy coming in as going out. We can approximate the Earth's loss of energy as if the Earth were a blackbody with temperature T_E, radiating into space. The blackbody range of frequencies for the Sun and the Earth are shown in figure 9.1. The energy loss is given by the Stefan–Boltzmann equation from chapter 2:

$$P_{\mathrm{out}}(T_E) = \sigma\varepsilon A T_E^4 = 5.67 \times 10^{-8}\ \mathrm{W\ m^{-2}K^{-4}} \times \varepsilon \times 4\pi r^2 \times T_E^4, \tag{9.2}$$

where $\sigma = 5.669 \times 10^{-8}\ \mathrm{W\ m^{-2}K^{-4}}$ is the Stefan–Boltzmann constant, A is area and ε is the *emissivity*, a unit-less constant which provides information about the physical properties of the emitting object. As mentioned in chapter 2, a perfect blackbody absorbs all frequencies, reflecting no energy, emits all frequencies and is defined to have an emissivity of one. In order to prevent the possibility of an object breaking the second law of thermodynamics by absorbing at a different rate than it emits, emissivity and absorptivity must be equal for a given surface in thermal equilibrium (Kirchhoff's law).

We can make a crude estimate of the Earth's surface temperature by equating the incoming energy from the Sun, acting as a blackbody with a surface temperature of 6000 K, with the outgoing energy of the Earth's surface, also acting as a blackbody but with a different temperature (Pierrehumbert 2010, Taylor 2005):

$$P_{\mathrm{in}} = (1 - a)S\pi r^2 = P_{\mathrm{out}}(T_E) = \sigma\varepsilon 4\pi r^2 T_E^4. \tag{9.3}$$

Assuming an emissivity of one gives a surface temperature of 254.8 K or −18.3 °C. The peak of the outgoing blackbody curve for this temperature is in the IR part of the electromagnetic spectrum. The Earth's average surface temperature is actually measured to be around 15 °C so clearly this calculation has omitted an important detail.

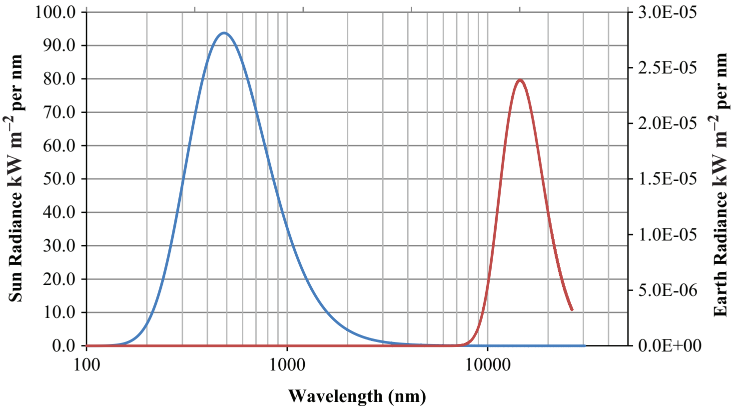


Figure 9.1. The blackbody radiance spectrum for 6000 K (representing the Sun, left) and 288 K (representing the Earth, right). Note that the scale for the energy output (radiance) of each is very different.

In 1863 it was discovered by John Tyndall that IR radiation is absorbed by water vapor and carbon dioxide, two trace components found in the atmosphere. It is the ability of the atmosphere to absorb energy that changes the surface temperature of the Earth to be in a habitable range. We can model the effect of the Earth's atmosphere by treating it as a third blackbody sandwiched in between the Sun and the Earth, absorbing and emitting energy as shown in figure 9.2.

Here we assume that the atmosphere is transparent to the blackbody radiation from the Sun because the peak of that radiation is in the visible part of the spectrum. Radiation from the Earth is in the IR range and a portion of this energy is absorbed by the Earth's atmosphere, as first shown by Tyndall. This layer eventually reaches equilibrium by radiating IR radiation both downward to the Earth and upward into space. The atmosphere acts in much the same way as building insulation does by slowing the rate at which heat is lost. All of the energy from the Earth's surface has to eventually be re-radiated by the atmosphere so we have

$$2\sigma\varepsilon_{\mathrm{a}}AT_{\mathrm{a}}^{4} = \sigma\varepsilon_{\mathrm{a}}AT_{\mathrm{E}}^{4}, \tag{9.4}$$

where the Earth's surface area A, at temperature T_{E}, is in contact with the atmosphere at temperature T_{a}. The emissivity of the atmosphere is used on both sides of the equation because only the part of the Earth's radiation which is being absorbed by the atmosphere is counted (reflection of visible light due to the albedo of 0.3 is not included in this energy balance). This result can be simplified to

$$T_{\mathrm{a}} = 2^{-1/4}T_{\mathrm{E}}. \tag{9.5}$$

Equation (9.3), the Earth's energy balance, can now be refined with an additional term due to the atmosphere. For an Earth with emissivity of one (since the reflected energy *is* being subtracted by accounting for the albedo in this equation) we have

$$\frac{(1-a)S}{4} + \sigma\varepsilon_{\mathrm{a}}T_{\mathrm{a}}^{4} = \sigma T_{\mathrm{E}}^{4}. \tag{9.6}$$

Substituting equation (9.5) into this equation gives

$$T_{\mathrm{E}}^{4} = \frac{(1-a)S}{4\sigma\left(\frac{1-\varepsilon_{\mathrm{a}}}{2}\right)}. \tag{9.7}$$

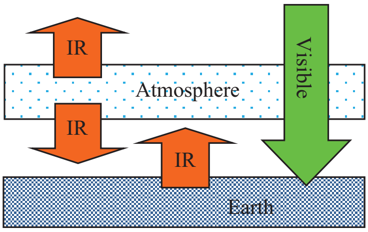


Figure 9.2. A simple model for the energy exchange between the Sun, Earth's surface and a single layer of atmosphere with a uniform temperature. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

A value of $\varepsilon_a = 1$ gives a surface temperature for the Earth of 303 K and a value of $\varepsilon_a = 0.75$ gives a surface temperature of 287 K, very close to the global average of 288 K. From this simple calculation it should be clear that the surface temperature of the Earth depends on the emissivity, ε_a, of the atmosphere, which in turn is a measure of the absorption properties of the chemical compounds making up the atmosphere.

As indicated by figure 9.3, the absorption/transmission properties of the atmosphere are wavelength dependent and are dictated by the chemical components of the atmosphere. From the figure it is clear that not all the energy coming from the Sun reaches the Earth's surface. There are many ranges of wavelengths which are blocked because of absorption by various molecules. As mentioned previously, the energy of a photon is $E = hf$, where h is Planck's constant and $f = c/\lambda$ is the frequency. If there are differences in electron energy levels in an atom or molecule corresponding to a particular photon energy, that photon can be absorbed by changing the electron's energy.

In addition to electron energy levels which can absorb photons of a certain frequency, molecules have quantized vibrational and rotational energy levels. The atmosphere is 78% nitrogen (N_2) and 21% oxygen (O_2) which are both diatomic and do not have energy levels in the IR photon range. Although they constitute the preponderance of the atmosphere they are largely transparent to visible and IR frequencies as a result of their chemical structure. Molecules of water, carbon dioxide and ozone have three atoms each and so have many more vibrational and rotational modes. In spite of their low abundance these more complex molecules have a much larger effect on absorption of electromagnetic energy from the Sun

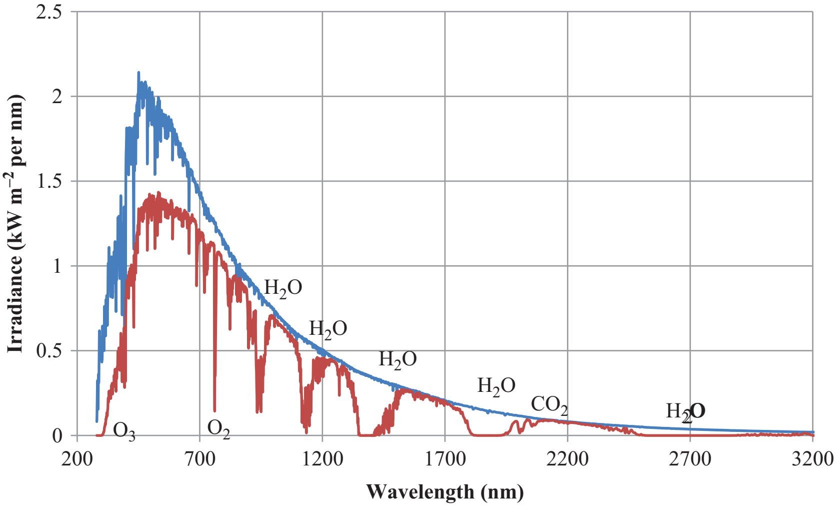


Figure 9.3. Energy reaching the top of the Earth's atmosphere (upper curve) compared to the energy reaching the Earth's surface (lower curve). Dips in the surface spectra are the result of absorption by different molecules and the responsible molecule is shown for larger dips (ASTM). From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

and in particular IR energy re-radiated by the Earth. Likewise methane, CH_4 and other more complicated molecules have even more vibrational and rotational energy levels with which to absorb photons and are even stronger absorbers (see Box and Box (2015) for details of atmospheric absorption, transmission and scattering).

In an isolated molecule the electron, rotational and vibrational energy levels would be very discrete (quantized) and absorption would occur only for a very narrow range of photon wavelengths. But because the atmosphere is relatively dense and at a warm temperature the molecules are constantly colliding, it has a range of thermal energy. Some of this extra kinetic energy can be 'borrowed' if a photon arrives with not quite enough energy to be absorbed. Likewise a slightly over energetic photon can be absorbed by a quantized energy level if it sheds some energy by adding to the thermal energy of the molecule. This broadening of the absorption spectrum increases with temperature and pressure. The absorbed energy is re-radiated (scattered in all directions) as part of the warming effect of the atmosphere.

So what happens to outgoing IR energy from the Earth? As seen in figure 9.1, the Earth's radiation occurs in a range of wavelengths that are much longer than those shown for the Sun's blackbody radiation shown in figure 9.3. The transmission of five atmospheric gases at a range of frequencies corresponding to IR emission from the Earth is shown in figure 9.4. The data shown are for these gases at the concentrations appearing in the atmosphere and at standard temperature and pressure. Water plays a significant role in the absorption of IR radiation, however, the broadest dip in transmission is due to carbon dioxide, which occurs at 15 000 nm—slightly to the right of the peak of the Earth's blackbody emission. The location and width of this blockage in transmission is why CO_2, a trace component

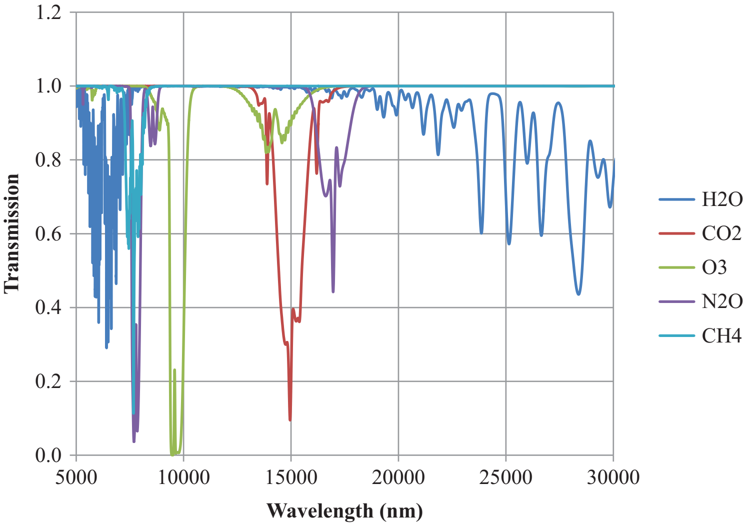


Figure 9.4. Atmospheric transmission spectra for several gases. A value of one means the gas does not block any radiation at that wavelength. The center of the Earth's blackbody emission spectra occurs near 10 000 nm (SpectralCalc 2012, Rothman *et al* 2009).

of the atmosphere, has such a large effect on the Earth's energy balance. As mentioned in chapter 1, stratospheric ozone (O_3) blocks harmful UV radiation from the Sun but as can be seen in figure 9.4 tropospheric, lower level ozone (primarily as the result of automobile emissions) ozone plays a large role in the absorption of IR from the Earth.

In the early 1900s it was argued that added carbon dioxide in the atmosphere past a saturation level could not block any additional outgoing IR radiation. This is not true, however, for two reasons. As the amount of CO_2 increases the dip in figure 9.4 does not get deeper but it will get broader due to the broadening of spectral absorption with concentration, a feature which could not be measured with the equipment available in 1900 (Pierrehumbert 2011). The second reason has to do with the fact that the concentration of CO_2 is not uniform in all layers of the atmosphere. If lower layers did become saturated, adding CO_2 can still have an effect by increasing absorption in higher layers which are not saturated.

There are many other dynamics of the global planetary system that affect the climate on Earth and other planets, for example, small changes in orbital parameters, ocean circulation patterns, small variations in the output of the Sun and various feedback mechanisms. However, by far the most significant for determining climate is the energy balance between incoming electromagnetic waves from the Sun and outgoing IR energy from the planet surface, modified by the atmosphere (Box and Box 2015). This effect is usually referred to as the greenhouse effect and the particular gases that are good IR absorbers are termed *greenhouse gases*.

The first step in making the model in equation (9.7) more accurate is to include a frequency dependent emissivity coefficient, based on the known absorption features of the molecules composing the atmosphere at their respective concentrations. The absorption properties change with temperature and atmospheric density so a further step is to break the atmosphere into layers, each with a different temperature and pressure. These calculations are most easily performed by computer and can be used to predict surface temperatures of other planets and planet satellites. For example the atmosphere on Venus is nearly pure carbon dioxide at 100 times the atmospheric pressure of the Earth. The spectral broadening of CO_2 absorption lines and subsequent IR absorption results in a surface temperature of 730 K compared to a temperature of 231 K if there were no atmosphere. Mars has an atmosphere 100 times thinner than the Earth and so has only a small greenhouse effect of about 10 K. For the temperatures and densities on Earth N_2 does not absorb IR, but on Jupiter's moon Titan, where temperatures are much cooler and the density of N_2 is higher, nitrogen does act as a greenhouse gas (Pierrehumbert 2011). The fact that radiation calculations (similar to those shown above) give reasonable surface and stratospheric temperatures for Mars, Venus, Earth, Titan and the upper layers of Jupiter, leads us to believe energy balance moderated by an atmosphere is the main constituent of planetary climate.

In order to be able to compare the effect of various gases on the surface temperature of a planet the terms *radiative efficiency* (in W m^{-2} ppb^{-1}) and *radiative forcing* (in W m^{-2}) are often used. Radiative forcing is the *extra* energy being absorbed by the planet's surface due to a particular gas given its abundance and

Table 9.1. Concentration of various components of the atmosphere and their radiative efficiencies (data from the IPCC). The amount of the radiative forcing due to human activity is noted as anthropomorphic.

Gas	Percent of the Earth's atmosphere	Radiative efficiency ($W\ m^{-2}ppb^{-1}$)	Radiative forcing ($W\ m^{-2}$)
Nitrogen (N_2)	78.1		~0
Oxygen (O_2)	20.9		~0
Argon (Ar)	0.93		~0
Water vapor	0.48		~56
Carbon dioxide (CO_2)	0.035	1.4×10^{-5}	~23 (1.66 anthropomorphic)
Liquid and solid water	0.002		
Neon (Ne)	0.0018		~0
Helium (He)	0.000 52		~0
Methane (CH_4)	0.000 17	3.7×10^{-4}	0.55 (anthropomorphic)
Krypton (Kr)	0.000 10		~0
Nitrous oxide (N_2O)	0.000 03	3.0×10^{-3}	0.15 (anthropomorphic)
Ozone (O_3)	0.000 007		0.14 (anthropomorphic)
Chlorofluorocarbons (CFCs)	0.000 000 14	0.01 to 1.4	0.30 (anthropomorphic)
Aerosols	0.000 000 02		−0.5 (anthropomorphic)

temperature in an atmosphere. Radiative efficiency is the extra energy being absorbed per mass of the molecule present per volume. So for example, methane has a radiative efficiency an order of magnitude larger than carbon dioxide, but because there is a lot more CO_2 in the atmosphere, the radiative forcing of carbon dioxide is much higher. Radiative forcing and efficiency for gases in the Earth's atmosphere are shown in table 9.1.

Water vapor is the predominant greenhouse gas in the atmosphere constituting 70% of the additional warming due to atmospheric gases. However the amount of water in the atmosphere changes in reaction to temperature—as the climate warms, more evaporation takes place and, conversely, if the atmosphere cools there is less water vapor and more water is found as ice. In this sense water vapor is a *feedback mechanism*, amplifying natural and human caused changes in climate. The hydrological cycle, which describes the movement of water from oceans into the atmosphere by evaporation from oceans and transpiration of plants and then back to the ocean as rain, runoff and ground water is well understood and is a self-regulating system, much less directly impacted by human activity than the carbon cycle (discussed below).

Water in the form of ice also acts as a feedback mechanism. Snow has an albedo of around 0.8 to 0.9 ice has a range of albedos between 0.5 and 0.7 but the open ocean albedo is around 0.06. If there is more polar ice there is a larger net

Earth albedo and less energy is absorbed, driving the climate into a colder regime. A warming climate produces less ice and more open ocean, which results in a much lower albedo, more absorbed energy and increasing temperatures.

The largest contribution to the Earth's albedo is cloud coverage, which reflects 20% of the incoming solar energy on average. The effect of clouds is complicated, however, and varies depending on the type and altitude of the cloud. High clouds are likely to have a warming effect by trapping more outgoing IR but the formation of low clouds likely has a cooling effect by increasing albedo. The Earth's daytime temperature rose slightly during the global grounding of airplanes after the terrorist attacks on the World Trade Center in 2001 because of fewer jet contrails reflecting solar energy. The nighttime temperatures dropped slightly, however, because more IR could escape and this is an indication of the complicated effects clouds have on surface temperature. It is the case that increased surface temperatures cause more ocean evaporation and increase the cloud coverage. However the extra moisture in the air also captures more IR radiation which tends to cancel the effects of a slightly increased albedo.

Aerosols are defined to be particles or droplets suspended in the air and small enough that they remain airborne for long periods of time. They may be solid or liquid and range in size between 1.0×10^{-5} m to 1.0×10^{-7} m in diameter. Rain will remove aerosols from the troposphere but if they reach the stratosphere aerosols may remain in circulation for as long as a decade. Natural sources of aerosols include volcanoes, dust blown into the air from deserts, forest and grass fires, and droplets from sea spray. Human sources primarily come from industrial process and burning, both from fossil fuels and to clear land for farming in developing regions. The effect of aerosols is generally to cool surface temperatures because they reflect incoming solar energy, but they can also scatter and absorb electromagnetic energy. For this reason aerosols are the least well understood part of climate[1]. An indirect effect of aerosols is they act as seed particles for raindrop and cloud formation (see Box and Box (2015)).

9.2 The carbon cycle

It was postulated as early as 1896 by the Swedish scientist Svante Arrhenius that adding carbon dioxide to the atmosphere might change the Earth's temperature (Weart 2016) but until Charles D Keeling started monitoring the atmosphere's carbon dioxide content it was not clear that human activities could affect the atmospheric carbon balance. Keeling started monitoring in 1958 at the atmosphere on top of Mauna Loa, Hawaii due to its isolation from local contamination, and he subsequently set up monitoring stations in other pristine locations such as Antarctica. As can be seen in figure 9.5, showing data from Mauna Loa, there is a an annual oscillation in carbon dioxide levels corresponding to plant growth in the

---

[1] Infrared energy is emitted from the earth's surface both during the day and at night with the result that atmospheric blocking of IR increases daytime and nighttime temperatures. Aerosols reflect daytime solar energy but have less effect at night. As a result of these two effects, measured nighttime average temperatures have increased more than daytime temperatures.

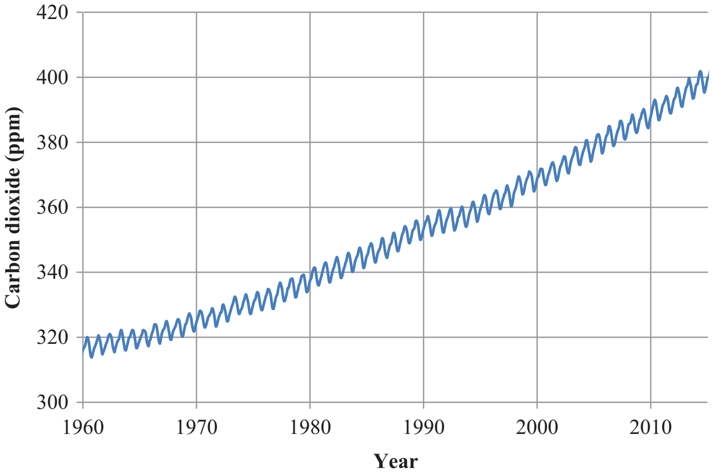


Figure 9.5. Carbon dioxide levels measured at Mauna Loa Hawaii: the Keeling curve (SIO 2016).

Northern Hemisphere in the summer and dormancy in the winter. Data from monitoring stations closer to the poles show less seasonal variation and the variation in the Southern Hemisphere is slightly less because there is less land mass where plants are growing. Although the annual contribution to the carbon flux due to humans is much less than the magnetude of the annual fluctuation of seasonal plant activity there is a cumulative effect causing the total atmospheric carbon dioxide to increase over time[2].

Because carbon dioxide has a disproportionate effect on the Earth's energy balance, a great deal of research has been done on the transport of carbon in the environment. The majority of the carbon cycle occurs due to natural processes, but a small portion is directly the result of human activity. Figure 9.6 is a summary of known carbon flux in the Earth's climate system for a recent year. The portion of the flux of carbon going into the atmosphere from human activities is small and mostly the result of burning fossil fuels: $C + O_2 \rightarrow CO_2 + \text{heat}$. Cement manufacturing involves the heating of calcium carbonate (limestone) to form calcium oxide (lime) which releases carbon dioxide, but this source, the second largest human caused CO_2 flux, is only 1% of the total global fossil fuel use emission. The land use change shown in figure 9.6 is an average with large regional variation and some parts of the Northern Hemisphere are currently carbon sinks (see Pielke *et al* (2016) for a review of the effects of land use change on climate). Land management practices, including reforestation, re-growth in farmlands, reduction of forest fires and better planting methods together with changing growing seasons and rainfall patterns may account for some of these regional differences.

---

[2] The large fluctuations in atmospheric carbon due to seasonal plant growth (100 billion tonnes per year) have motivated researchers to investigate genetically modifying the natural carbon dioxide fixation process in plants to be more efficient in turning CO_2 into usable plant fuels and to look for artificial CO_2 fixation pathways as a means to sequester atmospheric carbon dioxide (Schwander *et al* 2016).

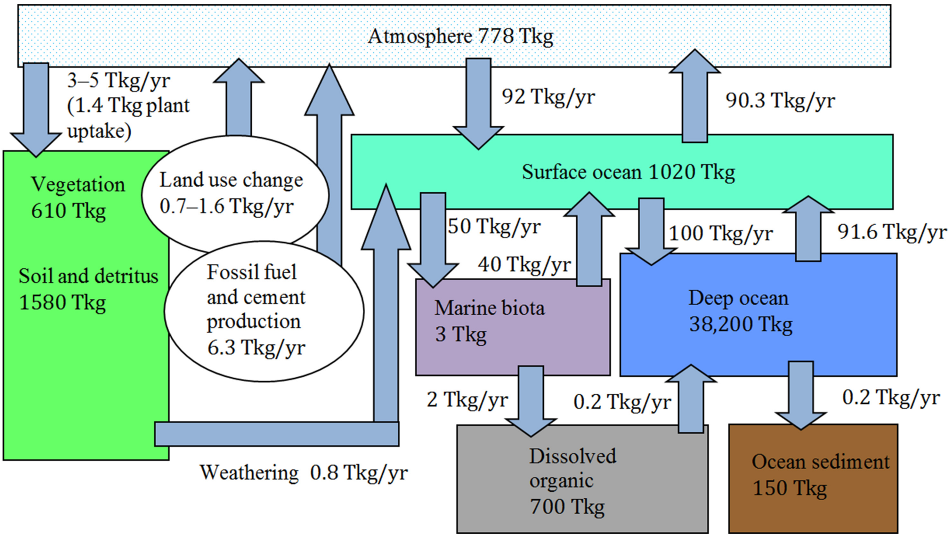


Figure 9.6. The carbon cycle in the year 2000. Arrows show carbon flux and boxes show carbon reserves in Tkg (10^{12} kg). Carbon in the ocean and atmosphere is in the form of carbon dioxide, carbon buried in ocean sediment or in soil may be in other compounds. Human caused changes in the carbon flux are shown as ovals (IPCC 2013, Graven 2016). From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

Several important points are shown in figure 9.6. The ocean, including the deep ocean, is the largest carbon reservoir. Ocean circulation patterns reduce carbon in the surface ocean by approximately 8.4 Tkg per year by moving this carbon to deeper layers. Since the turnover time for the thermohaline cycle is on the order of 1000 years, carbon transported to the deep ocean is effectively being removed from circulation at the present time. Currently the carbon flux between the atmosphere and the ocean is not balanced with the ocean absorbing about 2 Tkg per year which is about 30% of the extra carbon going into the atmosphere due to human activity. This extra carbon in the ocean is in the form of carbon dioxide, which is making the oceans more acidic, lowering the pH by 0.1 in the past 200 years. Marine biologists have established that higher ocean temperatures and acidity have caused several detrimental changes to various coral reefs around the world.

It is often useful to be able to compare the concentration of a substance in the air with the total amounts, as shown in figure 9.6. From the carbon in the atmosphere we can calculate the number of carbon atoms in the atmosphere which gives $778\ \text{Tkg} \times (6.02 \times 10^{23}\ \text{molecules/mole})/12\ \text{g/mole} = 3.9 \times 10^{40}$ molecules. The Earth's atmosphere at the surface is at a pressure of 10^5 Pa and since pressure equals force per area the mass of the atmosphere must be approximately $m = PA/g = 10^5\ \text{Pa} \times 4\pi \times (6.4 \times 10^6\ \text{m})^2/9.8\ \text{m s}^{-2} = 5.25 \times 10^{18}\ \text{kg}$. The atmosphere is mostly nitrogen and oxygen and we can obtain an approximate number of molecules of these two substances by using an average molecular weight of 29 g/mole. This results is $5.25 \times 10^{18}\ \text{kg} \times (6.02 \times 10^{23}\ \text{molecules/mole})/29\ \text{g/mole} = 1.10 \times 10^{44}$ molecules. The ratio of carbon to air molecules then is

$3.9 \times 10^{40}/1.10 \times 10^{44} = 0.000355$ or 355 ppm. This is approximately correct for the measured amount in 2000, which was 370 ppm. The current measured amount has recently reached 400 ppm, a 40% increase over pre-industrial carbon dioxide levels.

The added atmospheric carbon dioxide can be calculated based on the amount of fossil fuel being burned as energy resources. The ratio of carbon mass released to gasoline burned for fuel is about 3.3 kg CO_2 per kilogram of fuel, 2.7 kg CO_2 per kilogram for methane and averages about 3.6 kg CO_2 per kilogram of coal depending on the type of coal. Coal, which provides about a third of the fossil gasoline energy used globally, can vary from 95% to 50% in carbon content. A calculation of the amount of each of these fossil fuels burned in 2000 yields a number close to the measured 6.3 Tkg shown in figure 9.6, with cement manufacturing making up the difference, minus the amount that was absorbed by the ocean. This amounts to an additional 2.9 ppm of CO_2 in the atmosphere.

We can extrapolate backwards from the year 2000 to predict the pre-industrial carbon level assuming an exponential growth in fossil fuel use of 2.5% per year, comparable to population growth during this period. The decrease in total carbon dioxide from 2000 to 1800, based on a 3 ppm emission in 2000 is

$$C_{\mathrm{CO_2}} = \int_0^{-200} 3.0\mathrm{e}^{0.025t}\,\mathrm{d}t = \frac{3.0(\mathrm{e}^{-200} - \mathrm{e}^0)}{0.025} = -120 \text{ ppm}. \tag{9.8}$$

Subtracting this from the measured 370 ppm level in 2000 gives 250 ppm. From geological data, the average level of carbon dioxide in the atmosphere over the past 11 000 years was about 270 ppm. This type of calculation can be performed (more precisely) to extrapolate to future carbon levels depending on assumptions of the growth of different types of fossil fuel use.

Although the seasonal change shown in figure 9.6 is on the order of 8 ppm (53 Gkg of carbon flux) and the 6.3 Tkg added by human fossil fuel burning contributes less than 3 ppm, the cumulative effect since the Industrial Revolution has been a steady increase in atmospheric carbon. We now know that carbon dioxide levels for the past 800 000 years remained less than 280 ppm until the beginning of the industrial revolution 200 years ago (SIO 2016). This earlier data come from ice core samples, ocean core samples and other sources, discussed below. Figure 9.7 shows the *effective radiative forcing* (ERF) added to or subtracted from the solar constant (the zero line in the graph). ERF is the net radiative forcing of the source including the reaction of the system to the source. The anthropogenic sources shown are black carbon, jet contrails, land use changes and greenhouse gases (CO_2, methane, nitrous oxide, tropospheric ozone and CFCs). Natural sources of radiative forcing include aerosols, other aerosols and particles from volcanoes, stratospheric water, small changes in solar irradiance and stratospheric ozone.

As can be seen in figure 9.7, positive anthropomorphic radiative forcing far outweighs natural sources with the exception of short lived negative forcing from aerosols released from volcanoes. The net result represents an estimated 274 ZJ of extra energy added to the ocean/atmospheric system since 1970 (Box and Box 2015). Also seen in the figure, global temperatures generally drop by a degree or two for as

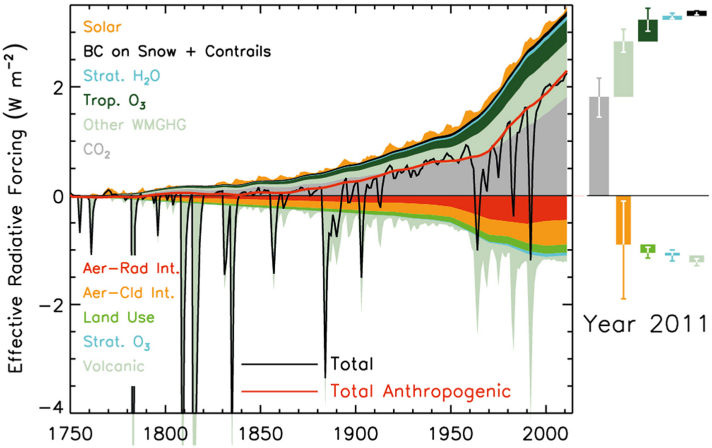


Figure 9.7. ERF over time of anthropogenic and natural causes relative to the solar constant (the zero line in the chart). BC stands for black carbon from fossil fuel sources deposited on snow. 'Other WMGHG' includes other greenhouse gases, such as methane, nitrous oxide and CFCs. 'Aer-' refers to aerosols which can interact directly to absorb or scatter radiation and in conjunction with clouds. The uncertainty bars represent 95% confidence. The total anthropogenic forcing is shown as a red curve and was 0.57 (0.29–0.85) W m^{-2} in 1950, 1.25 (0.64–1.86) W m^{-2} in 1980 and 2.29 (1.13–3.33) W m^{-2} in 2011. The total forcing, negative plus positive, is shown with a black curve (not to be confused with the black carbon curve) (IPCC 2013a).

much as several years after a large volcano erupts. The 22 year sunspot cycle causes the energy output from the Sun to fluctuate by as much as 1 W m^{-2} over the past 1000 years, but the effect is small relative to other forcings, as seen in the figure. Black carbon in the form of soot emitted from fossil fuel burning and volcanic activity, if deposited on snow, will reduce the energy being reflected and is thus a positive forcing. The figure does not show the forcing due to normal water vapor in the atmosphere since that is a feedback mechanism which adjusts to extant conditions, but the effect of stratospheric water is shown. Stratospheric ozone is a negative forcing but lower level tropospheric ozone traps IR as mentioned previously. The effect of aerosols from human activity is still not completely understood since they both scatter and absorb radiation, but it is know from volcanic aerosols the overall effect is a negative forcing since these substances reflect sunlight.

9.3 Paleoclimatology

Geologists have known for some time that the Earth's climate has changed radically on an approximately periodic schedule of hundreds of thousands of years known as the ice age cycle. The current ice age cycle started approximately 2.58 million years ago and has a periodicity of approximately 100 000 years. These changes represent a series of natural climate experiments over time that allow us to associate certain atmospheric conditions with surface temperature changes. Understanding these prehistoric climate changes allows us to build confidence in our understanding of how various factors affect the climate on Earth and to put the current sharp increase in carbon dioxide levels in context. A main source of information about the paleoclimate comes from various isotopes found in ice and ocean sediment core samples. These are cylindrical samples about 15 cm in diameter and sometimes hundreds of meters long, extracted using equipment similar to drills used in oil exploration. The later part of this record is corroborated with tree ring data and similar growth layers in corals. These sources of information are not direct measurements of climate but are referred to as *proxy data*. Other proxy sources include micro-organism fossils, prehistoric pollen deposits and coral growth rings. It is an important result that these proxy measurements agree where they overlap geographically and in time, as shown in figure 9.8.

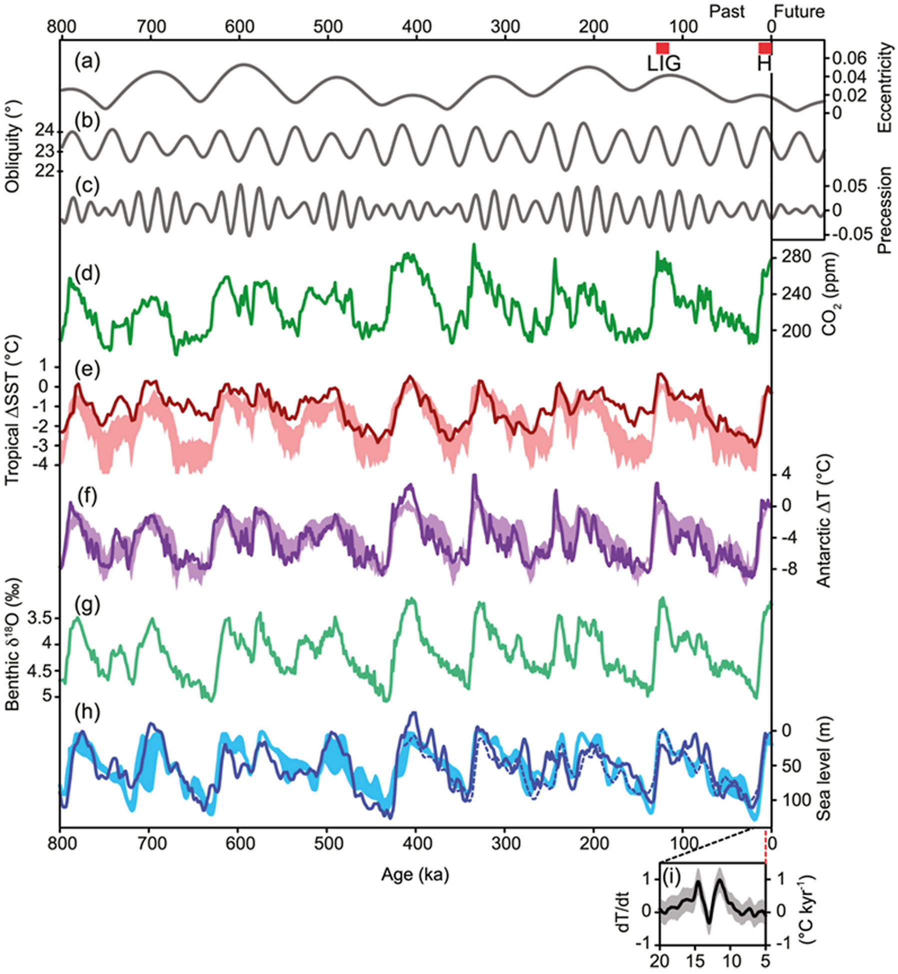


Figure 9.8. Climate proxies starting 800 000 years before the present. The top gray curves (a), (b), (c) are changes in the orbital parameters of the Earth. Curve (d) gives the carbon dioxide levels from ice core data (more recent data are shown in figure 9.5). (e) shows tropical sea surface temperature (SST) variations from the present day average. (f) shows Antarctic temperature deviation from the present average temperature based on ice core samples. (g) shows oxygen isotope levels at the sea floor, a proxy for ocean temperature and ice volume. (h) shows reconstructed sea levels (solid line) and computer models (shaded) (IPCC 2013b).

Dendrochronology is the use of tree ring data to establish information about ancient climatic conditions. New growth occurs in the spring and summer and the woody material added to the tree diameter is different than in the winter when the tree is dormant. The amount of growth is an indication of the growing conditions including temperature and rainfall. Isotopes of carbon vary slightly due to sunspot activity which changes the flux of cosmic rays arriving at the Earth, and tree samples have been used to determine sunspot activity over the past 11 000 years. Continuous records of overlapping tree ring data, including fossilized tree trunks, go back 11 000 years with some isolated spans going back much further. Samples have been collected from a wide selection of geographic locations across the globe.

Sclerochronology is the science of using coral and other durable marine organisms to learn about ancient ambient environmental conditions. Coral, mollusks and other sea life take up heavy metals such as strontium, uranium, thorium and magnesium at different rates depending on temperature and ambient concentration. As these organisms grow they form calcium compounds with these heavy metals in layers much like tree rings. Isotope data from these sources are one of the few direct way of establishing ocean levels over the past 600 000 years (Yokoyama and Esat 2016). Layers in stalagmites and stalactites can be used in a similar fashion to establish connections between ice age cycles and persistent local weather patterns (Wilson 2016).

In ocean and lake locations where the rate of *sedimentation* has been heavy there are often enough data to extract climate information from a core sediment sample. In a manner similar to tree rings, seasonal changes leave an annual record in lake and ocean sediments called a *varve*, which shows up as a different colored layer. Radiocarbon dating can be used to date the ages of the layers, and the amount and type of organic material give an indication of the climate the year the layer formed. Plankton and algae which live at the surface of the ocean or lake produce an organic compound called alkenone which comes in two different forms. Di-saturated alkenone acts as an antifreeze to keep the algae more flexible in cooler water. When the winter is harsh the ratio of di-alkenone to the warm temperature tri-alkenone form is higher. When these organisms die they leave a record of this ratio in the ocean sediment and measuring the ratio of these two types of alkenones in an ocean sediment layer gives an indication of sea surface temperatures at the time the organisms were alive. Pollen from land plants deposited in ocean sediment can also be a climate indicator. A preponderance of pollen from cold adapted plants indicates a cooler year.

There is a slight difference in the evaporation rate, called *fractionation*, of water where the oxygen in H_2O is the isotope ${}^{18}_{8}O$ compared to ${}^{16}_{8}O$. Water with the lighter isotope evaporates slightly more rapidly than water with the heavy isotope. The ratio of these two isotopes, therefore, is an indication of water temperature since warmer water has higher evaporation rates which changes the ratio. The microfossil remains of surface dwelling plankton in ocean or lake sediment retain a record of the oxygen isotope ratio of the water when the plankton died.

The longer the amount of time that water in the ocean spends away from the surface, the less dissolved oxygen will be present and sediment layers are affected by the amount of oxygen in the surrounding water. Less dissolved oxygen in an ocean sediment layer indicates a slower ocean turnover rate and hence a cooler climate when the layer formed. This can be statistically determined by the relative abundance of fossils of bottom dwelling organisms adapted to low oxygen conditions. It is also the case that some bottom dwelling marine bacteria adapt to low oxygen conditions by preferentially absorbing $^{14}_{7}\mathrm{N}$ compared to $^{15}_{7}\mathrm{N}$, so changes in the nitrogen isotope ratio map ocean floor oxygen levels and therefore ocean circulation.

The amount and type of inorganic material found in ocean sediment is also a clue to past climate and can be determined from the *magnetic susceptibility* (the degree to which it can be magnetized in an applied magnetic field) of the layer. Floating ice which originated from glaciers and polar ice caps has debris embedded on the underside as the result of sliding along land surfaces before being released into the sea. An increase in ice flow due to a warmer climate carries more and heavier debris further into the ocean, which can be found in ocean sediment layers. An ocean floor layer far from shore with a high density of inorganic material indicates a warmer year.

Fresh fallen snow is porous allowing a free air exchange with the environment. Over time as more snow accumulates, lower layers turn into a soft ice called *firn* which has trapped bubbles of air. If this ice occurs in a region where the ice does not melt in the summer it becomes a permanent record of the makeup of the atmosphere when the ice formed. Additional snow falling the next year creates a year by year layer of trapped bubbles. This process will build up layers each year, much like tree rings, which can be examined to determine atmospheric conditions in a given year. Annual variations from year to year mean that any sequence of multiple layers is unique. This makes it possible to compare cores taken at different locations and cores with slightly different time spans. If the time spans for two cores overlap they will share a unique sequence of layers and this allows the correlation of data from one core to that of another. Ice core data have been gathered from Greenland, Alaska, Antarctica, the Himalayas, Russia, China, the Andes and Mount Kilimanjaro in Africa. Some of the ice cores from these locations are hundreds of meters long, going back 800 000 years.

Atmospheric carbon dioxide, nitrogen, oxygen, methane and other atmospheric gas levels can be measured directly from the air trapped in bubbles. Temperature can be extracted based on the hydrogen isotope, deuterium, which has an extra neutron compared to normal hydrogen. This extra neutron means that water formed with deuterium is slightly heavier than normal water and ocean evaporation will favor the lighter form of water over water with deuterium when temperatures are lower. Deuterium is, then a proxy for temperature. This is similar to the oxygen isotope ratio mentioned above which also is an indication of evaporation rates and hence temperature. The $^{18}_{8}\mathrm{O}$ to $^{16}_{8}\mathrm{O}$ isotope ratio is another temperature proxy found in ice which can be used to corroborate the deuterium results.

Warmer climates result in larger regions of the ocean being ice free. More exposed ocean means more aerosols, dust and ocean salt are blown into the atmosphere and deposited on ice surfaces, changing the acidity of that layer of ice. The chemical changes in an ice layer due to these contaminates can be determined by measuring the electrical *conductivity*.

Sunlight forms hydrogen peroxide in gaseous form from photolysis of water vapor and measured levels of H_2O_2 in an ice layer are higher in years that have more sunlight. The rate of diffusion of a gas in an ice layer depends on the temperature difference between the ice layer and the surface and on the isotope of the element that is diffusing. The lighter ${}^{14}_{7}N$ isotope diffuses faster than the heavier ${}^{15}_{7}N$ isotope, which means that for any given layer, a larger separation difference between ${}^{14}_{7}N$ and ${}^{15}_{7}N$ indicates a warmer climate the previous year.

The beryllium isotope ${}^{10}_{4}Be$ forms in the upper atmosphere due to cosmic rays and the amount found in the atmosphere varies with changes in solar activity. The level of ${}^{10}_{4}Be$ in an ice core layer is therefore a record of sunspot activity, as well as changes in the Earth's magnetic field which block cosmic rays. Reversals in the Earth's magnetic field occur periodically over tens of thousands of years and can be independently determined from residual magnetic fields found in rock which has cooled after being exuded during sea floor spreading. When molten rock cools the iron in it becomes magnetized in the direction of the Earth's field. The further from the location where the molten rock is being deposited the longer ago the rock was formed and a clear record of magnetic reversals can be seen by examining rock further and further from the spreading zone.

Although the exact mechanism is not well understood, small changes in the Earth's orbit appear to be the root cause of the ice ages, as first postulated by Milutin Milankovic in the early 1900s. These orbital changes are due to gravitational interactions with other planets, particularly Jupiter due to its size. There are three types of perturbations: obliquity, eccentricity and precession, as shown in figure 9.8(a)–(c).

Obliquity is the tilt of the Earth's axis with respect to the plane of its orbit. Because of conservation of angular momentum the Earth's axis points in the same general direction as it travels around the Sun, much like a top or gyroscope. The North Pole is tilted towards the Sun for approximately six months and away from the Sun for the other six months when it moves to the other side of the Sun. When the Northern Hemisphere is tilted towards the Sun the Sun's energy strikes more directly causing summer, even though the Earth is in fact a bit further away from the Sun at this time of year. Currently the Earth's axis is tilted by an angle of about 23.4°, resulting in a seasonal variation of 400 W m^{-2} at northern latitudes, but this angle varies slowly between 22.1° and 24.5° over a 42 000 year period. Changes in obliquity do not change the annual total energy reaching the Earth, but these changes could affect the amount absorbed by as much as 6 W m^{-2} since there is currently more land mass in the Northern Hemisphere.

Eccentricity measures how elliptic the Earth's orbit is. The Earth's eccentricity varies from near zero (a perfect circle) to around 0.06 over a 96 600 year period and is currently about 0.017. At this eccentricity the solar constant at the Earth effectively changes by about 3% (40 W m^{-2}) as the Earth gets slightly closer and then further from the Sun. Averaged over the year, however, the same amount of energy reaches the Earth, regardless of changes in eccentricity.

Precession measures a wobble of the Earth's axis of rotation caused by gravitational interaction with the Sun and Moon on the Earth's equatorial bulge. The effect occurs over a 21 000 year period and was first noted by the ancient Greeks. Both precession and obliquity change the direction in which the axis of the Earth points, modifying which star will be the Pole Star. None of these three effects change the net annual solar energy reaching the Earth averaged over a year, but can affect the absorbed energy depending on which portions of the Earth are facing the Sun more directly.

Figure 9.8 clearly shows that diverse types of proxy data over the past 800 000 years change at the same times and rates (changes in the last 200 years are not shown because they would dwarf the prehistoric changes—CO_2 levels are now at 400 ppm, see figure 9.5). Other data show that methane and nitrous oxide follow the same pattern. Dips in the curves occur during ice ages when global average temperatures were on the order of 9 °C cooler and sea levels were more than 100 m lower than at present due to water frozen into ice sheets. We are currently living in an *interglacial* period that has lasted about 11 000 years, barely visible on the very right of the graph. The data also show that, although orbital parameters change slowly, the transition into and out of ice ages is rather abrupt. This is an indication that slow changes in insolation due to orbital variations are strongly amplified by feedback mechanisms such as reflection from ice coverage, changes in atmospheric water vapor and the amount of carbon dioxide in the atmosphere. It is the close correlation of CO_2 levels with temperature variations in paleoclimate data that supports the conclusion that trace amounts of atmospheric carbon dioxide have a large effect on global temperature as predicted by solar forcing calculations. The fact that the rate of change of temperature mirrors the rate of change of carbon dioxide concentration is the critical evidence that the two are closely connected. Data from times before the current ice age (3.3–3.0 million years ago and 50 million years ago) also indicate that very high levels of carbon dioxide (1000 ppm) are associated with temperatures higher than present day temperatures.

9.4 Climate modeling

The first computer climate models used radiation balance between layers of the atmosphere and known absorption properties and concentrations of atmospheric gases such as carbon dioxide to calculate the surface temperatures. As the models progressed in sophistication they added air and energy exchanges between large (500 km on each side) cubes of atmosphere on the surface of a flat Earth with differential heating to account for different angular orientations of the Earth's surface towards the Sun. By the 1980s clouds, ice and land masses were added. Eventually the models included the effects of rivers, aerosols, ocean and air circulation, atmospheric chemistry, seasonal effects of vegetation, and albedo variations due to land use changes. The cube sizes have also shrunk to 30 km, allowing for regional resolution of climate effects. There are some 23 independent computer models of varying sophistication run by more than a dozen different groups around the world.

The predictions of these computer models are not limited to surface temperature. Sea level changes, seasonal ice coverage, changes in stratospheric temperatures, rainfall patterns, drought patterns, frost patterns and growing seasons have all been calculated by computer models and verified with existing data for the recent past. The interested reader can download and experiment with a real climate model which will run on a desktop computer from the Educational Global Climate Modeling website (EdGCM 2016). This is a real climate model but with lower resolution and fewer features so that the results can be calculated in a reasonable period of time on a desktop computer.

Computer models are tested by averaging the output of several runs starting with slightly different initial conditions at dates in the recent past. In this sense they are very different from weather models which seek to predict specific weather events as compared to climate which is an expectation of the average condition for a particular region and time. These outputs are compared with the recorded average conditions to see if the models are accurate. The results of a series of comparisons are shown in figure 9.9.

As can be seen from figure 9.9, the computer models accurately predict surface temperatures for the past 250 years, but only if the anthropogenic forcings are included. As mentioned above, similar results are found for other output parameters such as rainfall, sea level and ice coverage, etc. The fact that climate models are found to be accurate should not really be surprising because they are based on well understood physical principles. The computer is simply performing the same types of calculations exemplified by equation (9.7), but including many more details which is not easily done by hand.

Predictions of future changes in climate depend on both current levels of CO_2 (which take on average 100 years to dissipate) and future human activity. Computer models can be run with various assumed human choices for the future, for example a freeze of all carbon emission at current levels or a continued expansion of fossil fuel use at the same rate as the past 200 years. The IPCC uses scenarios to label various assumptions about human future choices of energy sources. For the purposes of computer modeling these representative concentration pathways (RCPs) are given different amounts of estimated forcing. So, for example, scenario RCP2.6 assumes an additional anthropomorphic forcing of 2.6 $W\ m^{-2}$ over the natural forcing occurring in 1750, a figure close to the today's anthropomorphic forcing. The RCP2.6 scenario would be an assumption that humans do not continue to increase carbon emissions past what is being emitted currently. Figure 9.10 shows climate computer model predictions for surface temperature, ice cover and ocean pH.

9.5 Weather versus climate

There are many short term weather phenomena and it is important to be able to separate the temporary effects of weather from longer term climate trends. Varying weather features are generally related to the differential heating due to the angle of insolation and local albedo. The resulting air and ocean circulation patterns are modified by the location and shape of the continents. Because these temporary changes can be quite large and can vary on a time scale of as much as 30 years, it is sometimes difficult to distinguish them from longer term climate changes.

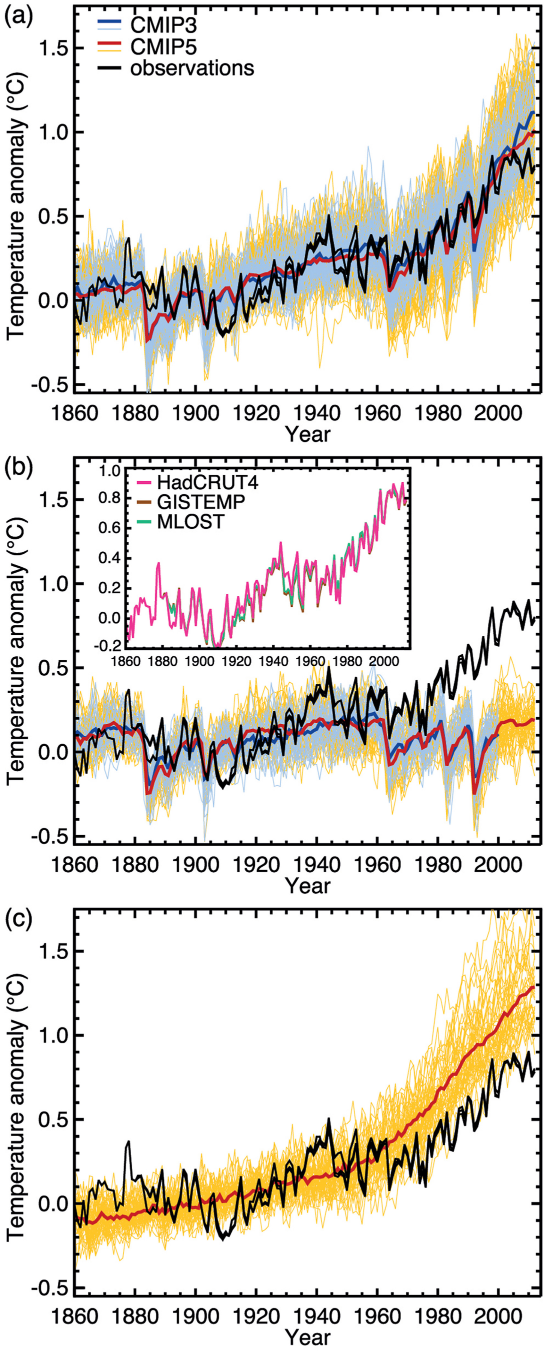


Figure 9.9. Averaged temperature data from three independent sources (black lines) compared to multiple runs on two different computer models (yellow and grey, with the red line the average of the yellow data and the blue line the average of the grey data). In (a) anthropogenic and natural forcings are included, (b) has natural forcings only and (c) shows only greenhouse gas forcing. The dips in 1963, 1982 and 1991 are the result of the volcanic eruptions of Agung, El Chichón and Pinatubo, respectively (IPCC 2013c).

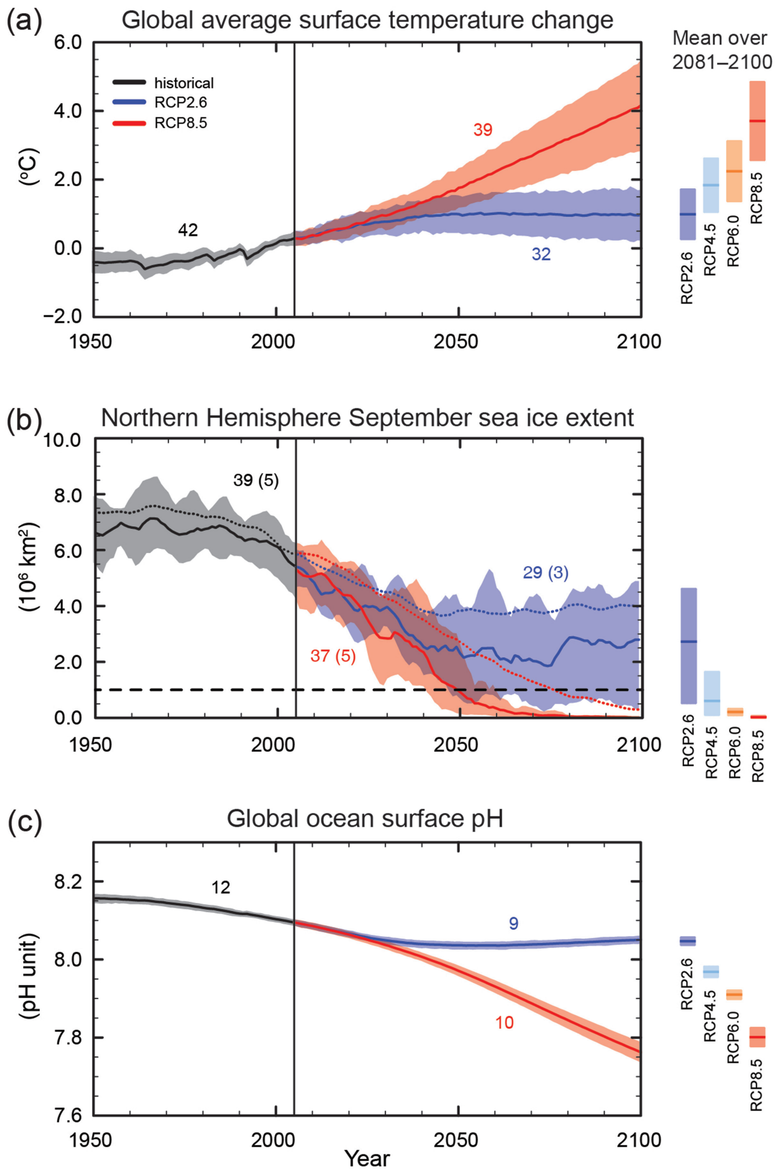


Figure 9.10. Model predictions of (a) average global surface temperature increases, (b) Northern Hemisphere ice coverage and (c) global ocean pH for various scenarios. RCPs are alterative scenarios of world economic development exemplifying different energy choices. The extra forcing for each scenario relative to 1750 is 2.6 W m^{-2} for RCP2.6, 4.5 W m^{-2} for RCP4.5, 6.0 W m^{-2} for RCP6.0 and 8.5 W m^{-2} for RCP8.5. The colored numbers refer to the number of computer models used to generate the curve with the matching color. (IPCC 2013d.)

A uniform planet which did not rotate but had an atmosphere would likely possess a very uniform circulation pattern. The Equator would preferentially heat because the angle of the incident solar radiation occurs at 90 degrees to the surface. The air in this region would heat and rise, spread radially alway from this region, sink at the Equator on the opposite side and flow back along the surface to the sunny side. A slow planet rotation would cause this circulation pattern to break into two with air rising along the Equator and sinking at the poles and this is approximately the case for Venus. The rotation also twists both the high and low airflows forming bands which are evident in IR photographs of Venus taken by satellite. The Earth rotates a bit faster than Venus so that the circulation pattern breaks into three circulation cells known as the Polar, Ferrel and Hadley cells. Air rises at the Equator and travels north at high altitude until it reaches 30 degrees latitude. There it sinks and flows south along the surface where it is deflected towards the west by the Earth's rotation due to the Coriolis effect. Between 30 degrees and 60 degrees the circulation of the Ferrel cell has air traveling north at the surface and this is deflected east due to the Earth's rotation. These winds, called trade winds and easterlies, respectively, made early trade between North America and Europe possible by sailing ships. A quasi-stable stream of air circulating the globe in the upper atmosphere at about 60° latitude called the jet stream is formed from the turbulence at the boundary between the Polar and Ferrel cells.

Warm surface water flows north at the ocean's surface from the Equator towards the poles where it sinks because it becomes cooler and saltier due to evaporation along the way. The cooler water returns to the Equator in a process called the *thermohaline circulation*. The entire trip takes about 1000 years and the rotation of the Earth, surface wind patterns, and the location and shape of the continents strongly affect the water circulation pattern. One familiar pattern is the Gulf Stream which carries warm water along the eastern coast of the US from the Caribbean to the North Atlantic. This circulation releases heat in the northern part of the North Atlantic and helps keep the climate of Northern Europe significantly warmer. Although Northern Europe is at the same latitude as Canada it has a much milder climate. Several rapid changes in prehistoric climate such as the Younger Dryas event, a sudden global cooling which occurred just as the last ice age ended, may have occurred due to a relatively rapid shift in ocean circulation patterns (Severinghaus and Brook 1999).

There are also several known regional oscillation patterns involving ocean and air circulation. The El Niño/La Niña Southern Oscillation (ENSO) has a periodicity of four to seven years and affects rainfall in the Peruvian desert, wind and rainfall patterns in the Pacific, and drought in Brazil, northeast South America and Central America. Under normal conditions, cool water upwells near the Equator along the coast of Peru and moves west towards the South Pacific. During an ENSO event, warm water in the Pacific Ocean at the Equator near Tahiti gradually spreads east towards the coast of South America, blocking the cooler water from surfacing. Indigenous people along the coast of Peru knew about the cycle long before the arrival of the Spanish because fish catch would be greatly reduced when the nutrient rich upwelling cooler water was not available to feed marine life.

Another oscillation pattern, the North Atlantic Oscillation (NOA), is a fluctuation between a semi-permanent low pressure atmospheric system which exists over Iceland and another semi-permanent high pressure zone in the Atlantic Ocean over the Azores off the coast of Portugal. When the low pressure over Iceland is lower than normal and the high in the Atlantic is higher than normal (a positive NOA), stronger storms are seen in the Atlantic, winters in Europe and North America are milder and wetter, while winters in Northern Canada and Greenland are colder and dryer.

The Arctic Oscillation and Pacific Decadal Oscillation are other ocean–atmosphere oscillations which occur with periods of 10–40 years. These various oscillations are referred to as internal variations in the climate literature and make the anthropogenic effects of climate change hard to discern over shorter periods of time. Figure 9.11 shows the relative magnitude of internal variations, solar fluctuations, anthropogenic effects and the effects of volcanoes, the largest contributors to climate forcing.

In figure 9.11 we can see that solar forcing is very small relative to other forcings and the effect shows a slight decrease over the past 50 years. Internal variation, for example the ENSO and NOA oscillations, although large, are not increasing over time but rather oscillate around a constant temperature. These rapid short term oscillations make weather more difficult to calculate than longer term changes in the climate which average out these faster oscillations. Only the anthropogenic forcing, shown in figure 9.11(e), has not fluctuated around a constant value and this increase explains the longer term temperature rise in figure 9.11(a).

9.6 Effects of human caused climate change

The basic argument that humans are causing a change in the climate can be stated in three claims, each of which has been verified by collecting substantial amounts of global data. (1) Humans are performing a massive experiment on the atmosphere, causing carbon dioxide levels to go from an average below 280 ppm over the past 800 000 years to a high of 400 ppm in the past 200 years. We know this is caused by human activity because the amount of fossil fuels burned during this time matches what is found in the atmosphere and ocean and, additionally, radiocarbon dating tells us this extra carbon comes from fossil sources. (2) We can calculate the effects of added atmospheric CO_2 on the Earth's surface temperature using straightforward laws of physics (blackbody radiation exchange and measured absorption features of the atmosphere). These calculations give the correct answer for surface temperatures on Earth currently and in the distant past, as corroborated by paleoclimatology, and for other planets and planet moons in the Solar System which have atmospheres. (3) There are measured changes in climate which correspond to the expected changes due to added CO_2. Figure 9.12 shows several climate indicators as they have changed over the last 150 years.

It should be emphasized that the charts in figure 9.12 are global indicators, collected by multiple teams of scientists. We can now say with certainty that global

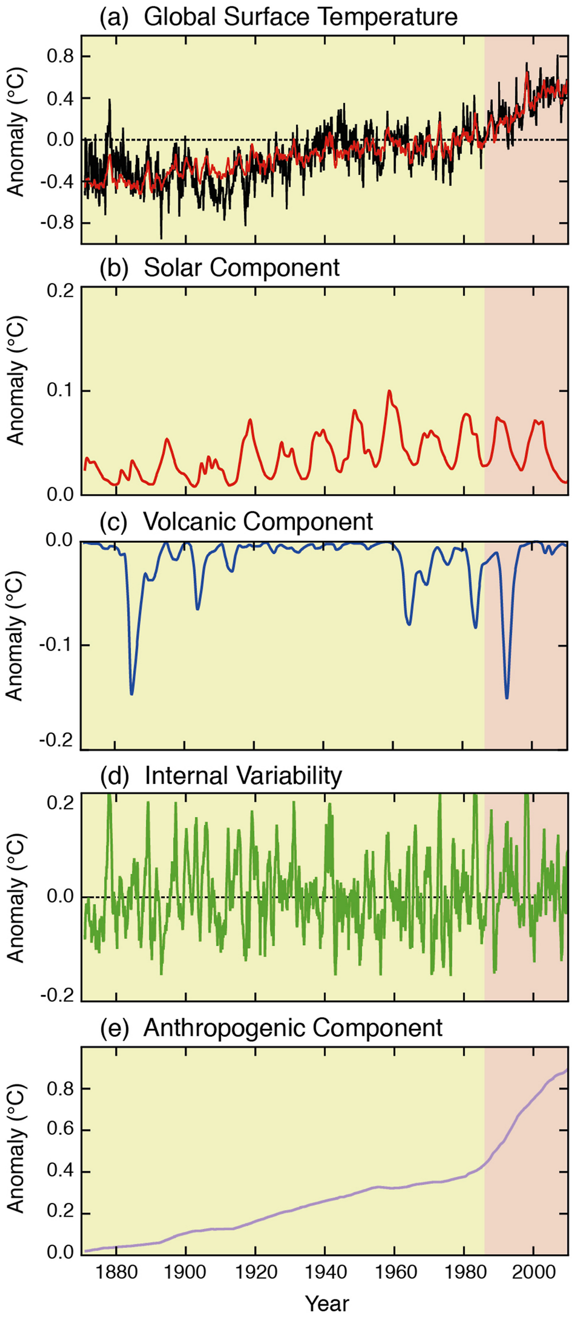


Figure 9.11. Global surface temperature record (1870–2010) relative to the average global surface temperature for 1961–1990 (black line) and a computer model of global surface temperature change (red line) produced using the sum of the impacts on temperature of natural, (b), (c), (d), and anthropogenic factors, (e). (b) Estimated temperature response to solar forcing. (c) Estimated temperature response to volcanic eruptions. (d) Estimated temperature variability due to internal variability, here related to the ENSO. (e) Estimated temperature response to anthropogenic forcing, consisting of a warming component from greenhouse gases, and a cooling component from most aerosols (IPCC 2013e).

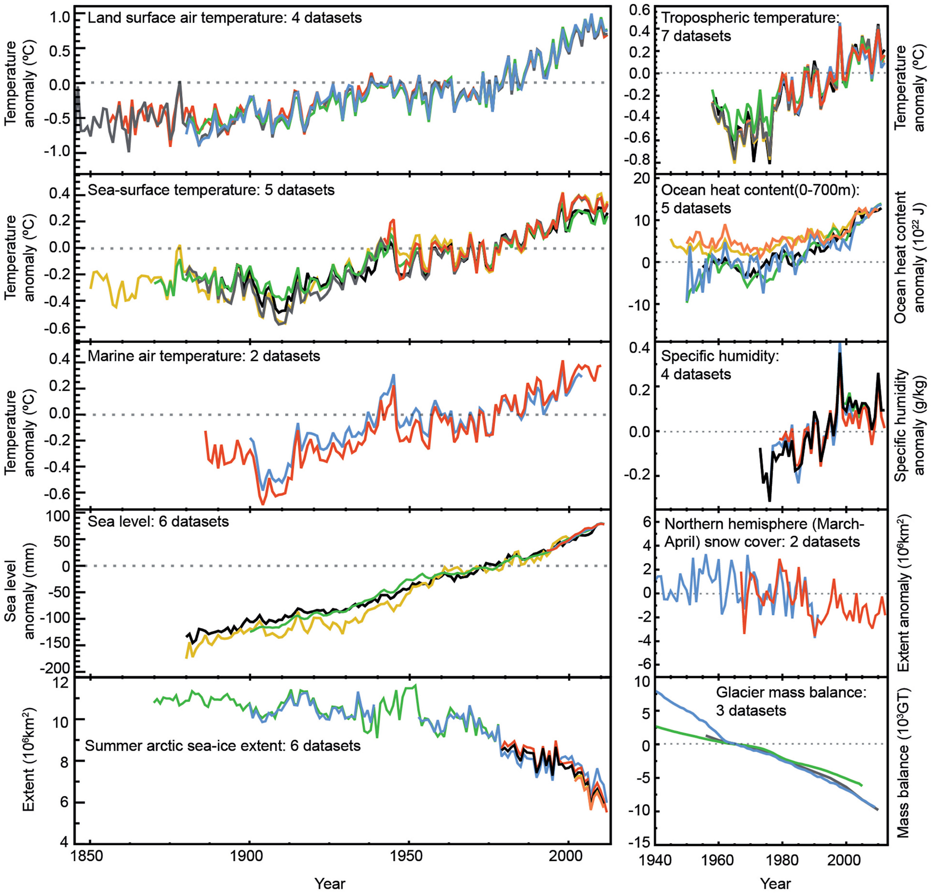


Figure 9.12. Multiple complementary indicators of a changing global climate. Each line represents an independently derived estimate of change in the climate element based on collected data.

temperatures have increased by 0.6 °C, sea levels have risen by 23 cm and arctic sea ice is shrinking (see, for example, NOAA 2015). Additional documentation shown includes shrinking glaciers, rising humidity, less ice coverage in the Northern Hemisphere and rising ocean heat content. The data shown in figure 9.12 are only a small part of the scientific data supporting the conclusion that global climate has changed in the past 150 years, roughly since the beginning of the industrial revolution (IPCC 2013).

One thing not shown in figure 9.12 is regional variations. Because the Earth radiates blackbody radiation nearly uniformly in all directions but absorbs energy from the Sun preferentially at the Equator, the IR flux from the Earth will be much larger than the incoming solar energy in polar regions. The result is a faster increase in polar temperatures than elsewhere, and this has been documented to be occurring. Changes in climate do not occur uniformly around the globe. Increasing ocean surface temperatures means more evaporation which leads to an increase in precipitation, but this effect varies regionally with some locations becoming dryer.

Increased ocean surface temperatures make more energy available to hurricanes and typhoons which are also regional effects. Current predictions of sea level rises are around one meter in the next 80 years which would affect a billion people living near ocean shores. Sea levels were 5–10 m higher than present during the last interglacial period.

Plants and animals are also reacting to changes in global climate. There are more than 28 000 documented cases of changes in biological marine and land systems including shifts in migration patterns, reproduction times, leaf maturation and plant blooming times (Rosenzweig *et al* 2008, Scheffers *et al* 2016, EPA 2016). Likewise the US Department of Agriculture (2012) plant hardiness zone maps, which tell farmers and home gardeners which horticultural plants will thrive in a given region, have shifted, in some locations by as much as 500 km. Further changes of this nature will have an impact on global food supplies. An estimation of the economic impact of a changing climate can be found in Nordhaus (2013).

Projects

1. The IPCC reports are available at http://ipcc.ch/index.htm. AR5 is the current full report. Other information is available such as technical and executive summaries. Pick two sections of the Executive Summary of the Technical Report and summarize them.
2. Real Climate www.realclimate.org/ is a website that tries to answer legitimate question about climate, run by real climate scientists. Write down five unanswered questions that you or your friends have after reading this chapter and look each one up on the Real Climate page. Discuss each question and the answers you find.
3. How to Talk to a Climate Skeptic http://grist.org/series/skeptics/ has responses to the most common skeptical arguments on global warming. Write down five questions or disputed facts that you or your friends have asked about climate change and look them up on this site. Discuss each question and the answers you find.
4. The Discovery of Global Warming www.aip.org/history/climate/index.htm is a thorough and accurate history, written by a historian of science, of the various scientific discoveries leading to the conclusion of human caused climate change. Read through the summary or some other interesting part of the book and comment on any surprising history that you did not know.
5. Download the EdGCM (Educational Global Climate Modeling) http://edgcm.columbia.edu/, a real climate model which can be run on a desktop computer. Parameters, such as CO_2 levels, can be changed and the simulation run for varying periods of time. Choose some parameters and see how they affect the climate over a 100 year period.
6. Write a report detailing the biological and/or economic impacts of climate change. Discuss both past impacts and likely future impacts.

References

ASTM (American Society for Testing and Materials) Reference solar spectral irradiance *Renewable Resource Data Center, National Renewable Energy Laboratory* http://rredc.nrel.gov/solar/spectra/am1.5/

Box M A and Box G P 2015 *Physics of Radiation and Climate* (Boca Raton, FL: CRC Press)

EdGCM (Educational Global Climate Modeling) 2016 http://edgcm.columbia.edu/ (Accessed: 11 July 2016)

EPA (US Environmental Protection Agency) 2016 Climate Change Indicators in the United States http://www.epa.gov/climate-indicators/downloads-indicators-report (Accessed 6 January 2017)

Graven H D 2016 The carbon cycle in a changing climate *Phys. Today* **69** 48

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

IPCC (International Panel on Climate Change) 2013 Climate Change 2013; the Physical Science Basis http://ipcc.ch/report/ar5/wg1/

IPCC (International Panel on Climate Change) 2013a Climate Change 2013; the Physical Science Basis (WG1AR5 figure 8-18) http://ipcc.ch/report/ar5/wg1/

IPCC (International Panel on Climate Change) 2013b Climate Change 2013; the Physical Science Basis (WG1AR5 figure 5-03) http://ipcc.ch/report/ar5/wg1/

IPCC (International Panel on Climate Change) 2013c Climate Change 2013; the Physical Science Basis (WG1AR5 figure TS-09) http://ipcc.ch/report/ar5/wg1/

IPCC (International Panel on Climate Change) 2013d Climate Change 2013; the Physical Science Basis (WG1AR5 figure SPM-07) http://ipcc.ch/report/ar5/wg1/

IPCC (International Panel on Climate Change) 2013e Climate Change 2013; the Physical Science Basis (WG1AR5 figure FAQ 5.1-1) http://ipcc.ch/report/ar5/wg1/

NOAA (National Ocean and Atmospheric Administration) 2015 Watch 25 years of ice disappear in 1 min http://www.youtube.com/watch?v=Fw7GfNR5PLA

Nordhaus W D 2013 *The Climate Casino* (New Haven, CT: Yale University Press)

Pielke R A, Mahmood R and McAlpine C 2016 Land's complex role in climate change *Phys. Today* **69** 40

Pierrehumbert R T 2010 *Principles of Planetary Climate* (Cambridge: Cambridge University Press)

Pierrehumbert R T 2011 Infrared radiation and planetary temperature *Phys. Today* **64** 33

Rosenzweig C *et al* 2008 Attributing physical and biological impacts to anthropogenic climate change *Nature* **453** 353

Rothman L S *et al* 2009 The HITRAN 2008 molecular spectroscopic database *J. Quant. Spectrosc. Radiat. Transfer* **110** 533–72

Scheffers B R *et al* 2016 The broad footprint of climate change from genes to biomes to people *Science* **354** doi: 10.1126/science.aaf7671

Schwander T, von Borzyskowski L S, Burgener S, Cortina N S and Erb T J 2016 Synthetic pathway for carbon dioxide fixation *in vitro Science* **354** 900

Severinghaus J P and Brook E J 1999 Abrupt climate change at the end of the last glacial period inferred from trapped air in polar ice *Science* **286** 930

SIO (Scripts Institution of Oceanography) 2016 The Keeling curve (real time graphs) https://scripps.ucsd.edu/programs/keelingcurve/

SpectralCalc 2012 High Resolution Spectral Modeling http://www.spectralcalc.com/info/about.php

Taylor F W 2005 *Elementary Climate Physics* (Oxford: Oxford University Press)

USDA (US Department of Agriculture) 2012 Plant Hardiness Zone maps http://www.ars.usda.gov/news-events/news/research-news/2012/usda-unveils-new-plant-hardiness-zone-map/

Weart S 2016 *The Discovery of Global Warming* (Cambridge, MA: Harvard University Press) http://www.aip.org/history/climate/index.htm

Wilson R M 2016 Paleoclimate record connects Asian monsoons and ice-age cycles *Phys. Today* **69** 13

Yokoyama Y and Esat T M 2016 Deep-sea corals feel the flow *Science* **354** 550

Chapter 10

Cost, benefit, risk

The two components of making decisions which involve the implementation of new technology are a *cost-benefit analysis* and a *risk assessment*[1]. The integration of these considerations with economic, social and legal constraints constitutes the focus of risk management. A cost-benefit analysis typically begins with a risk assessment and compares this with the economic cost of making some technological change assuming an estimated economic worth of protecting a human life or desirable resource. Often this is expressed as the cost for preventing a year of lost life (*cost per life-year*) due to a given risk. For example, the risk to drivers on rural roads involved in the placement of utility poles can be determined from historical data involving accidents with different arrangements of poles. The economic costs of either moving the poles further from the edge of the road or for using fewer poles can also be calculated. Comparing the risk with the costs allows the calculation of a cost per life saved for a given action and can determine the most economical way to decrease highway deaths due to the placement of utility poles (Wilson and Crouch 2001, Kammen and Hassenzal 1999).

Specific populations and situations may change the economic part of the evaluation. For example, if every newborn was screened for sickle cell anemia the cost would be roughly $34 million for every year of life saved (Wilson and Crouch 2001). However it is very unlikely that anyone without an ancestor from tropical regions of the world would have sickle cell anemia so it is not necessary to screen everyone. If only black newborns are screened the cost is about $240 per year of life saved. Economically speaking an expedient decision based on the dissimilar risks to different populations is to screen babies with heritage from tropical climates but not the general population.

---

[1] In economics risk assessment generally refers to economic risk. In this chapter the probabilistic risk of an event occurring is considered.

doi:10.1088/978-1-6817-4493-3ch10

10.1 Risk

A *probabilistic risk assessment* is an attempt to assign a quantitative value, either probabilities, lives lost, time lost, dollars, or other numerical indicators to a particular event or technology. The assessment may use historical data, for example actuarial statistics, and/or probabilities extrapolated from animal studies done in a laboratory (see chapter 1). In the case of analyzing risk from new technologies the failure rate of individual components may be combined to find the likelihood of a non-desirable outcome. Generally risk multiplies the probability of an event occurring times the severity of the event (Wilson and Crouch 2001, Kammen and Hassenzal 1999). So for example the risk of an experienced bicycle rider is lower than the risk of an inexperienced rider because the probability of an accident is higher for the inexperienced rider, while the severity is approximately the same. On the other hand, the risk for an experienced bicyclist not wearing a helmet is higher than for a rider with similar experience wearing a helmet because the severity is higher while the probability is roughly the same.

A risk analysis may be designed to answer the question, 'what is the probability that this event (for example exposure to radiation) will cause this result (for example cancer)?' (Kammen and Hassenzal 1999) for which the answer may be given in terms of the risk per mile traveled, per dose received or amount of exposure. Alternatively, the analysis may provide the risk for some occurrence from unknown causes either as a rate per year or risk accumulated over time. For example, the average annual cancer rate in the US for 2002–2006 for cancers from all causes was 542 per 100 000 for men and 409 per 100 000 for women, which represents a risk of 0.54% for men and 0.41% for women per year. Life expectancy in the US is 78.8 years for men and 81.3 years for women, which is a lifetime cancer rate of 43% for men and a 33% lifetime risk for women[2].

One useful way to report risk is the *relative risk*, which compares the risk for some group or population that has been exposed to a substance compared to an equivalent group which was not exposed. This is sometimes expressed as *excess relative risk*, which is the relative risk minus one.

A *probability of causation* analysis answers the question, 'was this outcome (cancer) caused by this event (exposure to radiation)?' (Wilson and Crouch 2001). Although the risk for certain types of cancer may be low due to exposure to some chemical agent or event, a person with that type of cancer may have a high probability of causation, indicating that there is a strong likelihood the cancer was caused by the exposure compared to other possible causes. The probability of causation for exposure to a given agent is calculated as the risk from that source divided by the risk due to all causes, whereas relative risk compares an exposed population to an unexposed group.

Everyone eventually dies from some cause but in a risk analysis we often are interested in the risk of *premature* death from exposure to some environmental

---

[2] This assumes the same annual risk for an entire lifetime which is not quite accurate because cancer rates are different at different ages.

Table 10.1. Days of lost life expectancy for a given risk (Wilson and Crouch 2001).

Risk	LOLE (days)
Alcoholic	4000
Smoker	2500
Cancer	1300
20% overweight	1100
Grade school dropout	950
Motor vehicle accidents	210
Air pollution	75
AIDs	55
Spouse smoking	52
Radon	45
Fire, burning	25
Poison	25
Nuclear Power	2

event. If death does not occur immeadiately, risk can be reported as *loss of life expectancy* (LOLE), *person days lost*, *working days lost*, *non-working days lost*, *quality adjusted life years* (QUALY) or *disability adjusted life years* (DALY). QUALY and DALY are often used to report risks due to unexpected events and exposures or diseases. It is frequently the case that these risk assessments apply to particular groups in certain locations, for example, workers in the nuclear power industry or the people of Flint Michigan (who experienced lead contamination as the result of changes to the water supply system). A short list of LOLE from various causes is shown in table 10.1.

QUALY, DALY and LOLE figures can also be used to report the risk of unexpected experiences such as a one-time exposure to a toxin or disease which may be debilitating but not immediately fatal. A few examples of risk calculation are given in this chapter but the reader is referred to the literature for a deeper discussion of these other measures of risk (see, for example, Wilson and Crouch 2001, Wilson 2002, Kammen and Hassenzal 1999 and Harte 1988).

10.2 Calculation of risk

An interesting, well understood example of a risk calculation is the risk of dying from a meteor impact. About once a year an asteroid the size of an automobile enters the Earth's atmosphere and once every few hundred years an asteroid between 50 m and 300 m in diameter strikes the Earth. The Tunguska event, an explosion occurring in 1908 over a remote region in Siberia, is now believed to have been caused by a large meteor impact. More recently there was a meteor impact in Chelybinsk, Russia in 2013 which injured around 1500 people.

The energy of an impact is easily calculated from the kinetic energy of the meteor: $\frac{1}{2}mv^2 = \frac{1}{2}\rho Vv^2 = \frac{1}{2}\rho\frac{4}{3}\pi r^3v^2$, where ρ is the density, r the radius and v the speed of the meteor[3]. Meteors entering the atmosphere at very shallow angles tend to be deflected back into space, but a solid meteoroid with a diameter of 50 m traveling at 20 km s^{-1} has a yield of about 10 MT of TNT equivalent and can penetrate into the lower Earth's atmosphere if it strikes at an angle near 90° to the surface[4]. A solid meteor with a diameter of 250 m traveling at the same speed would have a yield of around 1000 MT and likely leave a 5 km diameter crater on impact. The area of devastation from a large explosion in the atmosphere at an altitude of 25 km from the surface is given in km^2 by $A = 100\ Y^{2/3}$ where Y is the yield in megatons (MT) equivalent of TNT. In the area of impact most of the buildings would be flattened and the population in this zone would likely be killed more or less instantly. The 250 m radius meteor traveling at 20 km s^{-1} would, then, devastate an area of 1.0×10^4 km^2. Fatality rates based on the size of the impact can be predicted by assuming a population density for a given area.

The rate of impact and the size of the meteors striking the Earth has changed significantly in the 4.5 billion years since the formation of the Solar System. Before about 3.8 billion years ago asteroids produced large impact craters on all of the inner planets, including the Earth and the Earth's moon. Although much of the evidence of crater formation on the Earth has weathered away, more than 140 large impact craters have been identified, and the mass extinction of flora and fauna occurring at the end of the Cretaceous period has been linked to a meteor impact. The gravitational pull of Jupiter and Saturn, due to their large size, have gradually removed much of the small remnants left over after planet formation. The remaining asteroids have been striking the Earth, moon and other planets at a lower but more or less constant rate for the past 3.8×10^9 years. The rate of collision can be estimated by comparing cratering rates on the Moon and Mars, which have no or little atmosphere and Venus which has a thick atmosphere (Chapman and Morrison 1994). Direct observation in the form of meteor showers shows that some 25 tonnes of dust size particles enter the Earth's atmosphere each day. Table 10.2 gives meteor rates of impact, size of impact and an estimation of global deaths caused by a particular event. Very large scale events lead to a prediction of fewer deaths per year because they are so rare.

The known periodic meteor showers can be predicted because the orbits of the asteroids responsible for those showers have been identified. These objects are generally quite small and are not thought to be a threat. For larger objects NASA has a monitoring program that tracks those that have orbits close to the Earth's (NASA 2016). The Earth is not in any immediate danger, however the possibility of a collision with a large object is real. In 1989 asteroid 1989FC, with kinetic energy of 1000 MT, passed inside the Moon's orbit. Comets, although they are less dense than most asteroids and have orbits which make them less likely to strike the Earth, typically have speeds twice or three times as high as asteroids, making them significant threats as well. Collisions of comets with larger solar system objects do

---

[3] Typical meteor densities range from 1 g cm^{-3} to 8 g cm^{-3}.

[4] One tonne of TNT gives off approximately 4.18×10^9 J of thermal energy. The atomic bomb detonated over Hiroshima had a 15 kiloton TNT equivalent yield.

Table 10.2. Fatality rates for different types of meteor impacts (Chapman and Morrison 1994).

Event	Diameter (m)	Energy (MT)	Average interval (years)	Local deaths	Global deaths per year
Upper atmosphere	<50	<9	daily	0	0
Tunguska type	50–300	9–2000	250	5×10^3	20
Medium large	300–600	$2000–1.5 \times 10^4$	35×10^3	3×10^5	8
	300–1500	$2000–2.5 \times 10^5$	25×10^3	5×10^5	20
	300–5000	$2000–10^7$	25×10^3	1.2×10^6	45
Global	>600	1.5×10^4	7×10^4	1.5×10^9	2×10^4
	>1500	2×10^5	5×10^5	1.5×10^9	3×10^3
	>5000	10^7	6×10^6	1.5×10^9	250
Rare (Cretaceous type)	>10 000	10^8	10^8	5×10^9	50

occur and are sobering. Pictures of comet Shoemaker–Levy colliding with Jupiter in 1994 show spectacular, Earth sized holes in the surface layer of clouds on Jupiter.

An estimate of the risk of death by a global scale meteor impact for a person of lifespan 70 years can be made from the data in table 10.2. Using 1.5×10^9 deaths for an event that occurs once every 7×10^4 years gives a fatality rate of $1.5 \times 10^9/7.0 \times 10^4 = 2.1 \times 10^4$ deaths per year (shown in the final column of the table), about half the number of US highway deaths per year. Dividing this by the present global population of 7.4×10^9 gives a risk of one in 2.8×10^{-6} per year (a three in a million risk). This sounds like a small risk but a small risk over time continues to increase the probability of the event actually occurring. If you live for 70 years this would be a lifetime risk of $70\ \text{yr} \times 2.8 \times 10^{-6} = 2.0 \times 10^{-4}$ or 200 in a million, similar to the lifetime risk of being hit by a car as a pedestrian.

The risk from meteors is an interesting example of the risk from a low probability event which has a very high consequence or severity. Tunguska type events only occur every 250 years and an event that is large enough to kill off a significant fraction (20%) of the global population occurs only once in 70 000 years. Our perception of the danger of a meteor impact is low because we have no direct experience with such a disaster, unlike tornadoes or hurricanes. Statistically, however, although there are no documented cases of a human dying from a meteor impact, the risk of death from a meteor is roughly the same as dying in an airplane crash or flood[5].

10.3 Compared risks

Risk assessments are often put to use to compare the risk between two different scenarios in order to make a decision about a particular course of action, regardless of cost (Wilson and Crouch 2001). The conclusions, however depend on exactly how the risk is calculated. One approach is to find the *absolute risk* of each action and compare them. For example, suppose the risk for illness related to air pollution is

---

[5] There are, however, several known cases of automobiles and houses being struck by small meteors.

initially X per person in a city and different types of incinerator plants are proposed. Plant A is known to cause an incremental increase in risk of pollution related illness of 15%, but the plant is small enough that it can be located so that only a third of the population is affected. The incremental increase in risk for plant B is only 5%, but it must be located in a site that will affect half the population. Which plant should be chosen? The risk before the plant is build is NX, where N is the total population. The absolute risk for the people affected by building plant A is $N/3 \times 1.15X = 0.38\,NX$ and the absolute risk of the people affected by plant B is $N/2 \times 1.05X = 0.52\,NX$. This would suggest that plant A is the better choice because the absolute risk is lower.

A different approach is to calculate and compare the *incremental risk*, which is the increase in risk of each of two different actions. The incremental risk for plant A is $(N/3) \times 0.15X = 0.05\,NX$ and the incremental risk for plant B is $(N/2) \times 0.05X = 0.02\,NX$. Here we see that the *incremental* increase in risk due to plant B would be lower.

In the previous two calculations we have not included the unaffected population. Perhaps a better evaluation is the shared or *total* risk to the entire population due to either course of action. The total risk for plant A is $(N/3) \times 1.15X + (2/3) \times 1X = 1.05\,NX$ and for plant B we have $(N/2) \times 1.05X + (N/2) \times 1X = 1.02\,NX$, a 3% difference. Our conclusion from this evaluation is that plant B is the lower risk.

A decision as to which plant should be built cannot be made solely on the basis of a risk calculation because political, social and economic factors will be involved. Based solely on a risk evaluation, the two populations in the previous example will undoubtedly have different opinions as to which plant should be built. The severity of a particular risk must also be evaluated. If the pollution related illnesses in the example are life threatening or debilitating the decision becomes more difficult than if the effects are slight. The difference in cost of the two plants is also an important factor. Suppose plant B in the example costs 10 times as much as plant A, and the severity of the health problems are known to be very mild. How do we decide if the additional cost warrants the reduction in total risk by 3%?

10.4 Risk of new technology

The previous examples of risk calculations were based on known historical data but this method cannot be used to calculate risk of new technology. Although nuclear reactors have existed for more than 50 years the failure rate is too low to have statistically significant death rate data. In these cases the probability of failure of a complex system can sometimes be calculated based on the likelihood of failure of the individual components. A *fault tree* or *event tree* analysis is applicable for new technology if something is known about the failure rate of the individual components making up a complex system. For example the failure rate of cooling pipes in a nuclear power plant can be estimated from historical experience with similar pipes in other types of plants.

As explained in chapter 5, nuclear reactors cannot explode as an atomic bomb because the fuel is not sufficiently enriched to have a high enough component of the uranium 235 isotope. It is possible, however, for them to release radioactive material to the environment in the event of a serious disaster, such as a core meltdown (Fukushima) or graphite fire (Chernobyl). For this reason nuclear reactors (and most chemical plants and other dangerous manufacturing processes) are built with multiple levels of safety mechanisms designed to prevent a serious disaster in the event of mechanical or electrical failure. Figure 10.1 shows a simplified version of a fault tree for a reactor with the following branches: (a) rupture of a supply pipe (probability P_r); (b) loss of electric power (probability P_e); (c) failure of the emergency cooling system (probability P_c); and (d) failure of containment vessel (probability P_v). The probability of each of these *not* occurring is one minus the probability of it happening

If the probabilities for each component in a fault tree are known, the probability or risk of an event occurring can be calculated. Suppose the probability for each of the stages (rupture of a supply pipe, loss of electric power, failure of the emergency cooling system and failure of containment vessel) in figure 10.1 is known to be 0.1 (10%) per year based on historical failure rates of similar components. All of the possible scenarios ends in the successful containment of emissions except the bottom one. For the single failure mode we have $P_e \times P_r \times P_c \times P_v = 0.0001$ per reactor year or a one in 10 000 chance per reactor per year. A real life example of an event tree branch leading to a disaster is the failure of the Fukushima Daiichi reactor in Japan after an earthquake and tsunami in 2011. The reactor shut down as designed and cooling pumps began to cool the reactor, which takes several days until the core is safe from a possible meltdown. Unfortunately the diesel engines running the cooling pumps were protected from the tsunami by a wall designed to withstand a 5.7 m tsunami, a once in 100 year event. The actual tsunami was estimated to be 13 m to 15 m tall, a one in 1000 year event. The core melted down and the intense heat produced hydrogen which exploded, which introduced radioactivity into the atmosphere and ground water. In contrast, the Chernobyl disaster was primarily the result of human actions as many safety features were turned off by the operators, changing the fault tree and making any risk analysis useless (Cohen 1987).

The risk of failure of reactor components is not the same thing as the risk of exposure of the population to radioactive material. In the Three Mile Island incident

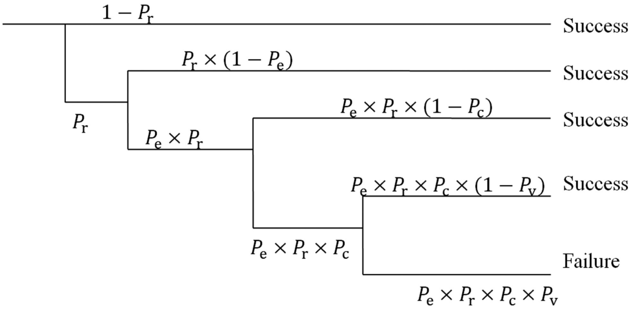


Figure 10.1. Extremely simplistic fault tree analysis of for an emission of radioactive gas from a nuclear reactor due to a loss of coolant. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

in 1979 in Pennsylvania, the most serious reactor accident in the US to date, there was a coolant system failure due to a failed valve which led to a core meltdown, but there was no hazardous exposure to the public. Deaths from the Chernobyl disaster are estimated at 49 immediate deaths and some 4000 premature deaths due to thyroid cancer. The magnitude nine Tohoku earthquake and subsequent tsunami at Fukushima killed approximately 16 000 people and dislocated a quarter of a million more in Japan, but there were no lives lost due to the reactor meltdown.

As a final point a fault tree analysis can sometimes be corroborated by calculating an upper bound of the likelihood of failure of the system based on absence of historical failure. For example before the Fukushima reactor disaster there were more than 8000 reactor years of operation worldwide for light water reactors (the most common type of commercial reactor) without a loss of cooling accident (the Chernobyl reactor was not a light water reactor). This puts an upper bound on the likelihood of a loss of cooling accident of $1/8000 = 1.25 \times 10^{-4}$ per reactor year for light water reactors. The estimate based on fault tree analysis is an order of magnitude lower than this.

10.5 Perception of risk

Humans are notoriously inaccurate at intuiting risk. As noted in chapter 8, riding in a car is a very risky proposition compared to most other forms of transportation. Traveling 300 miles in a car carries about the same risk of a fatal car crash (one in a million) as traveling 1000 miles in a plane but many people would rather drive the 1000 miles (and face three times the risk) than take a plane. A single chest x-ray carries the same fatal cancer risk (one in million) as living two months in Denver[6] and both of these carry the same risk as living 150 years in a 20 mile radius of a nuclear power plant, yet the general public has a strong aversion to nuclear power plants. Table 10.3 lists the major causes of death in the US, given as deaths per 100 000.

Various factors affect our ability to accurately evaluate risk. For example our evaluation of risk is strongly affected by exceptional events. Airplane crashes are rarer and have a much lower risk per mile than auto crashes but plane crashes kill a large number of people all at once. Not a single death occurred in a commercial plane accident in 2002 compared to 42 000 automobile deaths in the US, but many people view flying as more dangerous. Nuclear reactor failures or plane crashes are more easily remembered than car crashes which causes us to vastly overestimate the risk of flying and nuclear power. A recent, highly publicized failure of a product, for example a tire or seat belt, may cause us to ignore the true safety record of the product. If a failure occurs early rather than later in the production lifetime of a product the perceived risk of the brand is much higher. For example, suppose a particular model of airplane has a landing gear failure rate of once in 100 000 landings. If the failure occurs in the first landing rather than in the 100 000th landing the perceived risk of that model airplane will be much higher although the probability of failure is the same.

---

[6] Denver has an altitude of 1610 m and so has less atmospheric protection from cosmic rays.

Table 10.3. Risk of death due to an action or cause in the US (Wilson and Crouch 2001, Kammen and Hassenzal 1999).

Cause	Annual per capita risk per 100 000
Heart disease	271
Diabetes	24
Pneumonia and influenza	35
Cigarette smoking (per smoker)	300
All fatal cancers	200
Motor vehicle accident (total)[a]	11
Alcohol related	6
Using cell phone[b]	0.2
Home accidents (all ages)	11
Illegal drugs	5.6
Alcohol	6.3
Falls	6
Over age 70	43
Suicides	11
Accidental poisoning	4
Yearly coast to coast flight	1
Being hit by falling aircraft	0.004
Being hit by a meteorite	0.04
Tornado	0.015
Flood	0.045
Lightning	0.016
Heat wave	0.4
Bites or stings from venomous insect	0.017
Radon	
US average of 1.5 pCi l^{-1}	0.015
US high of 400 pCi l^{-1}	0.045

[a] These figures are about twice as high for the US as other developed countries (CDC 2015).

[b] See NSC (2016) for more details.

In cases where we can easily imagine the probabilities for something to fail we tend to evaluate the risk as much higher than if we cannot think of a way for the failure to occur. For example, a fault tree analysis for a manufacturing plant may list many ways failure can occur which paradoxically can convince someone that the process is inherently dangerous in cases where the calculated possibility of failure is quite low, specifically because the system has a complex set of safety features.

Humans are strongly influenced by social norms and this changes our risk perception. Probably very few people plan to drive while under the influence of alcohol or drugs, but the social milieu of a party influences us to take what is actually a very high risk. Socially in the US it is considered quite normal to make a quick shopping trip to a store when in fact this is one of the more risky things we do on a day to day basis.

Personal choice and control play a large role in our evaluation of risk. Many people are willing to talk on the phone while driving, but very few of us would choose to experience a tornado close up although the risk is an order of magnitude lower for tornadoes. Some people are afraid of flying because they do not have control of the plane, but in fact pilots are very highly trained and much more experienced than the average driver of an automobile. The risks for many popular recreational activities (for example all-terrain vehicles) are much higher than many occupational risks, but because we choose to do them we discount the amount of risk involved.

Latency or the delay between cause and effect makes it very difficult for humans to accurately evaluate risk. Many smokers discount the risk of smoking because they are not yet ill. Likewise being arrested for a single illegal drug infraction is relatively unlikely which may sway a decision to use illegal drugs.

The subjective evaluation of dread and familiarity with a technology plays a large role in risk perception. Radiation from any source is often evaluated as much more dangerous than second hand smoke from cigarettes because of subjective dread of radiation and familiarity with cigarettes. Radiation cancer risks from nuclear power plants are much lower than cancer related to many other chemical exposures, yet most people perceive nuclear power plants to be more dangerous. Nearly 3000 people died in the 11 September 2001 attack on the World Trade Center, giving people a strong perception of terrorism as a serious threat to life, whereas annual influenza, which kills between 12 000 and 50 000 people a year in the US (and is preventable with a vaccine), is often not perceived as a serious danger.

High risks to small groups are typically not perceived to be as serious as smaller risks to large numbers of people. The seriousness of food allergies has come to the attention of food manufactures and the general public only recently because the risk was restricted to a smaller portion of the population. Because it had the potential to affect a large number of people, saccharine, which has been proven to be relatively safe, received much more public attention than exposure to tree nuts, which affects a smaller number of people, although this smaller number is affected much more severely.

The perception of risk is lower if people trust the person or persons or agency in charge of that risk. Knowledge of this fact is the reason why the tobacco industry continues to donate large sums of money to charity in an attempt to sway the public perception of risk introduced by tobacco products. Conversely, perceived risk rises sharply if it is discovered that officials have misled the public as occurred in the Flint Michigan water contamination crisis.

Figure 1.1, which shows the death rate by age group for the past 100 years in the US, highlights several other features of risk perception. As can be seen in the figure, the influenza epidemic of 1918 was a much larger factor in early deaths than either of the World Wars. There is a measurable increase in death rates among the 15 to 24 year old group starting in the late 1960s due to the rapid rise of the number of cars available and a sharp increase in young drivers. A slight increase in death rates for 25 to 44 year olds in the mid-1980s reflects the AIDs epidemic. Infant mortality has dropped precipitously since 1900, largely as the result of widespread vaccination and the use of antibiotics, but many people today incorrectly think vaccines constitute an unacceptable risk. Probably none of these noticeable features of death statistics would be in the average person's lists of most significant risks to life.

10.6 Economic factors

As noted previously, scientific assessment of risk is, unfortunately, seldom definitive in decisions about the application of given technology. Economic, social and political factors usually play a significant role in making a decision regarding environmental threats. Deciding to ban CFCs was relatively easy because the economic impact was comparatively small, since replacement materials were known and already under development. Banning carbon dioxide emissions is much more problematic because 80% of the energy supply for modern society is implicated. The economic costs of switching to renewables will be significant, even if the cost of renewables is competitive with existing fossil fuel sources and the long term costs are lower.

According to modern economic theory, as the price of a commodity being sold increases the demand decreases. On the other hand, as prices increase sellers will want to increase the number of units sold and so supply will increase. Economists define an *efficient market* to be the case when demand matches supply and this occurs for a specific price[7]. If the price is too low or demand increases there will be a *shortage*, an incentive for sellers to increase supply until efficiency is reached. If the price is too high or demand decreases there will be a *surplus* and sellers will be motivated to decrease supply, again arriving at equilibrium. This perfect balance is rarely achieved in practice because it assumes fixed costs for production, raw material and labor over time as well as unchanging technology. It also assumes consumers are not affected by changes in preferences (i.e. advertising, new information about the potential harm or benefit of a product, etc), income, cost of living or other variations in personal circumstances and they have perfect knowledge about prices offered by competing sellers. These factors all affect the buyer's willingness to buy a product[8].

The seller is affected by costs and potential profit. Costs for a seller involve the fixed costs of building a manufacturing plant or establishing a service, fixed costs of maintenance and operating the production process or service, as well as the variable costs of raw material, unexpected maintenance and operation costs, for example labor. As we saw in table 6.3, in the energy sector fixed costs include the *capital cost* of building the power plant, normal maintenance and operation of the plant, as well as variable costs of fuel and variable maintenance and operation costs of upgrading and labor. Once a production process is operating sellers are also concerned about *marginal costs* (the cost to use the existing production process to produce one more unit of a commodity or service—an additional MWh for example) and *marginal*

---

[7] Probably the most widely used definition of efficiency is the *Pareto* efficiency which is defined to occur when any change in the market (prices, supply or demand) makes the public worse off than before (where 'worse off' and 'better off' are relative terms).

[8] The assumption that all other factors remain constant while surplus, demand and prices fluctuate is called the condition of *ceteris paribus*.

revenue (the additional revenue from the additional unit produced). For electricity generation the marginal cost of making more electricity from an existing coal plant is much lower than the cost of building a new, more efficient plant.

The cost of a given technology generally decreases over time and its energy efficiency will increase as the technology improves. As an example of this trend, the cost of a kilowatt hour of electricity generated by solar cells has decreased at a rate of about 20% per year over the past 20 years, as has the cost of wind energy. As a technology matures the technological gains over time decrease. Gas turbines are a mature technology and as a result have been decreasing in cost at a rate of 10% per year, which is a much smaller change than solar or wind technology. Cost is also heavily influenced by *economies of scale*; constructing a factory to build many windmills will result in a lower cost per windmill compared to building them individually.

One additional factor that plays a large role in the adoption of any new technology is the *time value of money*. If P is the present value of a resource (for example investment dollars) and F is the future value the relation between the two is given by $F = P\mathrm{e}^{rt}$ where r is the interest rate and t is time. The inverse of this equation tells us that \$10 paid out 20 years in the future is worth $P = F\mathrm{e}^{-rt} = \$10\ \mathrm{e}^{-0.1\times 20} = \$1.35$ in today's money if the available interest rate is a constant 10%. Another, somewhat counterintuitive way to view this is that future dollars are worth less than current dollars because if we have that dollar today we can invest it in such a way that it will be worth more in the future. In this case the rate of change is called the *discount rate* instead of the interest rate. Conventional business planning typically has relatively short time horizons because the value of money and interest rates change over time, making it more difficult to predict the return on a longer term investment.

The changing value of money is important in the energy industry because it is necessary to borrow money to build power plants and these plants are often in operation for up to 60 years (see figure 10.2). The interest rate charged for borrowing the money and the change in value of the money over time must both be factored into the final cost of the electricity produced. The *capital recovery factor* (crf) calculates what the annual payment will be in today's dollars if we borrow P dollars at an interest rate of r for t years:

$$\mathrm{crf} = \frac{r(1+r)^t}{(1+r)^t - 1}. \tag{10.1}$$

The annual payment, A, is

$$A = P \times \mathrm{crf} \tag{10.2}$$

so, for example, if we borrow \$20 000 for 15 years at 6% interest our annual payment will be $A = \$20\,000 \times \frac{0.06(1+0.06)^{15}}{(1+0.06)^{15}-1} = \$2059.26$. This factor takes into account the time dimension of the value of borrowed money. Likewise a payment in the future must be 'discounted' to know what it is worth in present day dollars. Other useful economic analysis factors (which we will not discuss further in this book) include

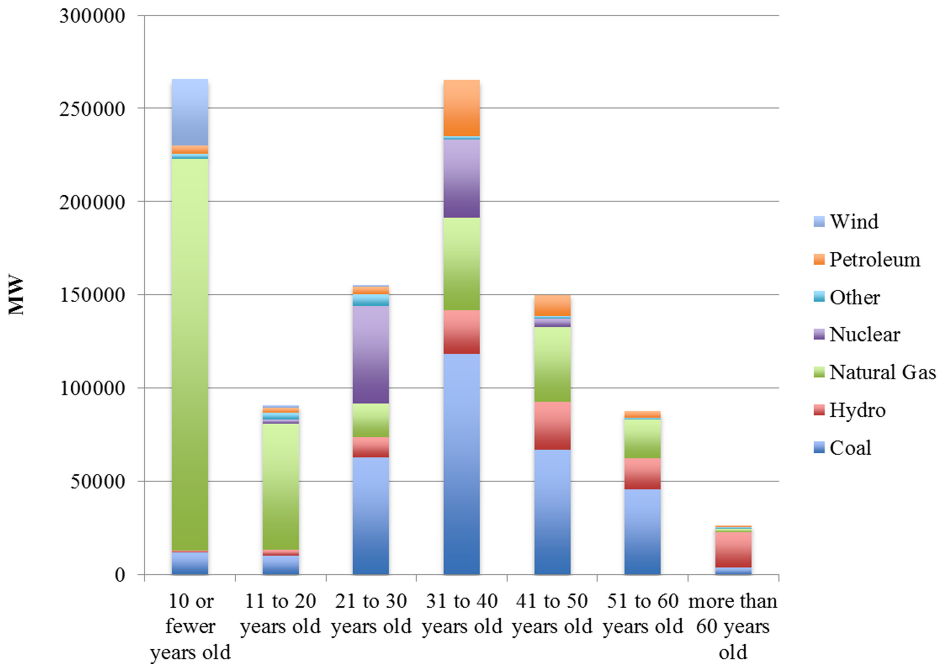


Figure 10.2. Electric power generation in the US in 2010 by source and age of plant. The increase in natural gas plants over the past ten years reflects a drop in the cost of natural gas as the result of hydraulic fracturing.

the *compound growth factor* (used to find the future value of a current investment at a given interest rate), the *present value factor* (used to find the value of future dollars in terms of present dollars), and the *uniform present value factor* (used to compare annual payments to a one time payment given the discount rate of future dollars). These concepts were used to calculate the levelized cost (the yearly revenue needed to recover all expenses over the lifetime of the investment) shown in table 6.3 for electricity generated from various renewable resources.

The decision as to whether investing in a particular technology will be profitable or not is often based on a calculation of the *net present value* (NPV). The NPV is calculated as the sum of cash income minus costs (including initial costs) during a series of time periods (usually a year) where the future costs and incomes are adjusted using the change in value over time (present value factor). If the NPV is positive (for a given assumed interest rate) the investment will make money, if NPV is negative it will lose money.

The economic factors related to environmental issues are complex, particularly for the economics of climate mitigation. The reader is referred to the references for more in-depth discussions and examples of economic factors in the energy industry and for climate concerns. See Crets *et al* (2007), Deutch and Lester (2004), NAS (2010), Randolph and Masters (2008), Tester *et al* (2005), Nordhaus (2013) and Stern (2007).

Most economic interactions involve costs which neither the seller nor the buyer pays for directly, and these are referred to as external costs or *externalities* and these may be positive or negative. A typical example of a positive externality is having your child vaccinated for a disease which benefits other children because your child cannot transmit that disease to other children who are not vaccinated. The parents of the other children do not pay for the benefit to their children. A classic example of a negative externality is pollution generated by an electric power plant. Life-cycle greenhouse gas emissions of various electricity generating power plants are shown in figure 10.3, and table 1.1 shows other pollutants from various energy sources. The cost of ameliorating the effects of these poisons is not included in the direct cost to the seller or consumer of the electricity produced. The lack of an economic incentive to protect a free, commonly held resource such as the atmosphere or a water resource is referred to as the *tragedy of the commons.*

Greenhouse gas emissions are not the only external costs of electricity generation and both traditional fossil fuel resources and renewable energy sources have external costs. For example, building a hydroelectric dam, often considered the cleanest source of electricity, involves environmental and societal disruptions in building the dam as well as the emission of methane and carbon dioxide from organic material accumulating behind the dam. Coal power plants give off sulfur compounds, toxic organic compounds, as well as heavy metals such as mercury, all of which are harmful to the environment and the organisms (including humans) that absorb these pollutants. Life lost in coal mining disasters, although much lower in modern times, is another external cost not directly paid for by the power industry. All electricity sources require land which cannot be used for other purposes and the location of a power plant changes the property value of surrounding acreage. Solid waste in the form of ash from coal plants and spent fuel rods for nuclear plants require disposal or, for nuclear waste, long term storage (a problem which still does not have a satisfactory solution). The water required for cooling of conventional generation sources leaves the water at a much higher temperature than when it enters the plant

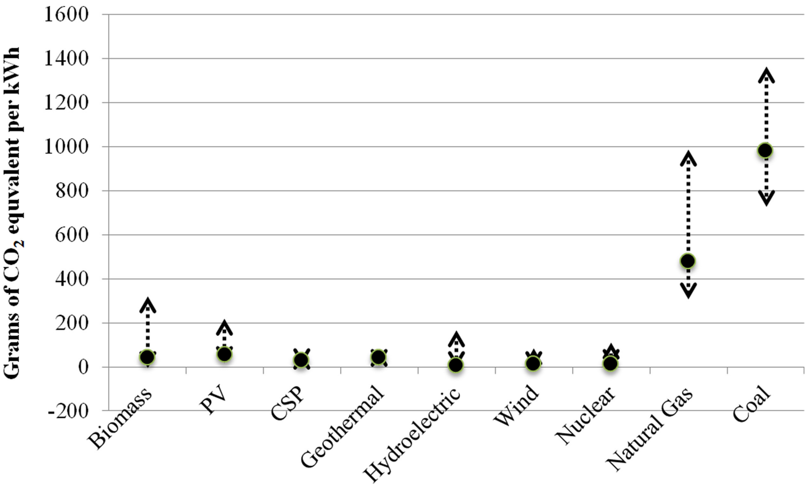


Figure 10.3. Life-cycle greenhouse gas emissions of several types of electricity generation for equivalent amounts of energy generated. CSP is concentrated solar power (thermal solar) and PV is photovoltaic. The data are an average made by the US National Renewable Energy Laboratory of several hundred life-cycle analysis reports. The dotted lines give the range of the reported data (NREL 2016).

and this can cause damage to marine environments. Residents living near windmills sometimes complain of noise and the visual impact of the windmills which constitutes a local negative externality. Petroleum has an energy security externality for the US in the sense that the military presence and diplomatic efforts used to guarantee access to resources in other regions of the world are paid for by the government (via tax payers) and not the energy industry (Crets *et al* 2007, Roth and Ambs 2004).

It is sometimes possible to put a price on an externality once it has been identified. For example the external cost of electricity generation from renewables has been estimated in the US to be around 2 cents per kWh while coal and natural gas have external costs of approximately 10 cents per kWh, based on considerations of the cost of several of the individual externalities listed in the previous paragraph (Energy Institute 2017, Roth and Ambs 2004)[9]. The EPA in the US quantifies the economic damage occurring from direct and climate related damage due to carbon emission, known as the *social cost of carbon* (SCC) at $56 per metric tonne of CO_2 emitted assuming a 5% discount rate (EPA 2015)[10]. A calculation of the SCC involves making an assumption about the future cost of economic damage (risks) as the result of deteriorating environmental conditions compared to the cost of present day investments to fix them. Nearly all economists agree that making investments now to prevent future costs of climate change is economically favorable in the long term but the outcomes are very sensitive to the choice of discount rate[11]. Various models (so called integrated assessment models (IAMs)) which calculate the economic results of spending money today to mitigate climate related economic problems in the future are available online. One such model is the Dynamic Integrated Climate Model (DICE) by W D Nordhaus (Nordhaus 2013).

Other external costs are more difficult to evaluate, particularly if they are positive externalities such as free access to environmental resources like parks and recreational lakes. A society may decide it is worth preserving a location, for example a National Park, from developers who wish to harvest timber, mine coal or drill for oil in that area. It might also be desirable to preserve rainforests, waterways, marine environments or exotic species in other locales, even though the benefits might not be directly accessible to everyone. To accomplish these goals requires that a price be paid for these services which may involve paying taxes if they are to be preserved as a public good, available to all. These external costs can sometimes be determined by a process of evaluating the components of the service or product, for example the number of people who might use a national park, the number of scenic views, campsites, species being protected, etc[12]. For example, a value is affixed to each of

---

[9] Estimated external costs for electricity generation are higher in Europe (EEA 2007).

[10] At the time of this writing the new administration of the US is looking at changing the formula for calculating the SCC.

[11] Actually it is the *social discount rate* which is in dispute. The social discount rate is the discount rate on projects which have a (difficult to measure) future social impact such as infrastructure, education and climate change mitigation.

[12] This process is referred to as a *hedonic regression* and is used to determine the Consumer Price Index in the US, among other things. The Consumer Price Index uses the cost of a chosen set of agreed upon desirable goods to evaluate the effects of inflation, changes in wages, pensions and other economic factors affecting consumers.

the components based on what the average person might pay which would allow the comparison of value between two National Parks that offered different amounts of similar services (see Dasgupta and Ehrlich (2013) for other types of social externalities).

A government has several ways to 'internalize' negative external costs into the price, for example, of electricity from a particular source, which typically cannot be internalized by an open market. Control based legislation mandates internalization of cost by regulations. The mandating of scrubbers or other technology for cleaning smokestack emissions of coal fired plants is an example as are the CAFE standards for car efficiencies in the US. This also requires setting up a system for monitoring plant emissions to make sure the maximum allowed discharge does not occur. The cost of the scrubbers is paid for by the operator of the plant which passes the cost on to the consumer in the form of a price increase in electricity.

A second way to internalize external costs involves taxation or incentive based legislation which can take several forms. As mentioned in chapter 1, the problem of acid rain in the US was greatly improved by an artificial market created to trade taxes on sulfur emissions. This system of *cap and trade* uses market incentive to drive down the amount of emissions by placing a price on pollution and also encourages the development and implementation of new technologies to further reduce emissions.

In some cases demand for renewable energy may be increased by direct marketing to consumers, even though the cost is slightly higher. As noted in table 10.4, the US government spends a significant amount of money to encourage end users to use energy sources wisely. One goal of these inducements is to create a market for renewable energy which eventually will allow economies of scale and innovation to reduce the cost so that they become competitive with traditional energy sources. Incentives can be phased out once a new technology has matured. Unfortunately the policies supporting these types of incentives often change too quickly to be effective, often on a scale of two to three years. There were large fluctuations in the early implementation of wind power in the US in large part due to short term instabilities in tax credits being offered by the government. Several countries in the European Union, however, have guaranteed minimum prices for renewable energy for periods as long as ten years. For example France has fixed a price of 0.082 euros kWh^{-1} for renewable energy for ten years (National Academy of Science 2010). This approach, called a feed-in tariff, levels the playing field for a significant amount of time allowing energy suppliers and consumers to make long term plans and investments.

Reducing taxes and providing other financial incentives for companies developing new technologies, sometimes called process based legislation, can also encourage the creation and improvement of new processes that have lower externalities. Table 10.4 shows the range of incentives, including tax breaks, given to various sectors of the energy industry in 2013 in the US. Loan guarantees given by the Department of Energy were available before 2013 but are no longer used to incentivize the US energy sector.

A simple example can demonstrate the effects of a rebate and tax incentive on the cost of electricity. Suppose a home owner wants to install a rooftop PV system which

Table 10.4. US government subsidies to the energy sector in millions of dollars. Direct costs are subsidies given directly to an industry. Tax costs are tax incentives. RUS is the Rural Utility Service provided by the Agricultural Department. ARRA is money provided by the American Recovery Act of 2009. 'End use' is money spend on efforts to incentivize consumer reduction in consumption. LIHEAP is the Low Income Energy Assistance Program. The total subsidy is around $90 billion (EIA 2015, NRC 2010).

Beneficiary 2013	Direct	Tax	Research and development	Federal and RUS electricity	Total	ARRA related
Coal	74	769	202	30	1075	129
Refined coal	–	10	–	–	10	–
Natural gas and petroleum liquids	62	2250	34	–	2346	4
Nuclear	37	1109	406	109	1660	29
Biomass	332	46	251	–	629	369
Geothermal	312	31	2	–	345	312
Hydropower	197	17	10	171	395	216
Solar	2969	2076	284	–	5328	3137
Wind	4274	1614	49	–	5936	4334
Other	209	–	380	5	594	229
Biofuels	72	1670	74	–	1816	6
Smart grid and transmission	8	211	831	134	1184	780
Conservation	833	630	501	–	1964	1574
End use	3513	1997	466	–	5976	2046
LIHEAP	3116	–	–	–	3116	–
Other	397	1997	466	–	2860	2046

costs $15 000 and generates 3000 kWh yr^{-1} (2000 W peak output). If they take out a 20 year loan at 6% the crf is 0.0872 and the annual payment comes out to be $1307.77 which is $0.44/kWh, much higher than the current average of $0.12/kWh in the US but competitive with average electricity prices in the EU. Suppose, however, there is a rebate from the state for renewable energy of $2.00/W of energy generated and a 30% federal tax credit on the cost of the panels after the rebate. In this case the rebate reduces the capital cost by 2000 W × $2.00/W = $4000, so it is only necessary to borrow $11 000. The federal tax credit of 30% reduces the capital cost by $3300 which reduces the annual payment to $671.44 which cuts the cost of electricity in half to $0.22/kWh.

Most of the fossil fuel industry is strongly opposed to carbon taxes, although most economists believe this is the most effective way to decrease greenhouse gases (Nordhaus 2013)[13]. A metric tonne of coal, of which 70% is carbon, provides about 8200 kWh of thermal energy. A 35% efficient coal plant would generate 2900 kWh

---

[13] Economic analysis in (Nordhaus 2013) indicates that a carbon tax may be the *only* way to limit global warming to 2 °C as regulations are not effective enough to produce the required reductions in CO_2 emissions.

of electrical energy at this efficiency. Suppose a \$100 per tonne carbon tax is imposed. This represents a \$0.034/kWh increase in the cost of electricity; a 28% increase over the cost in the US and an 8% increase over average electricity cost in Europe. Although the cost will inevitably be passed along to the consumer there are several very strong arguments for using a carbon tax to capture externalities related to carbon dioxide emissions. A carbon tax evenly distributes the burden across the entire energy sector and has the additional advantage of being far less complicated than regulation on individual energy sectors. This broad based approach also incentivizes the energy industry to find better solutions to emissions, for example developing better carbon capture and sequestration (CCS) techniques or switching to less carbon intense sources. Currently it costs about \$100/tonne to sequester carbon, but this price is expected to decrease to \$20/tonne by 2025, which would make it cheaper to sequester carbon than to pay a proposed carbon tax of \$25/tonne.

One additional factor that must be included in a detailed cost analysis of energy resources is the capacity factor, which is the ratio of actual power output to the rated capacity of the plant. Fossil fuel and nuclear power plants can be kept online as much as 90% of the time with maintenance and other offline periods making up only 10% of the lifetime of the plant. As noted in table 6.3, plant capacity factors for most renewable sources are generally only rated at 20% to 50% because of their intermittent nature. This means that a windmill rated at 4000 kWh yr^{-1} with a capacity factor of 25% can only be relied on to supply 1000 kWh yr^{-1}. To supply a demand of 4000 kWh yr^{-1} would require a plant with a peak rating of four times this number (or four such plants) coupled with a storage system with no losses.

10.7 Energy efficiency and the economy

Tax incentives, low interest loans for development, consumer education programs as well as mandated efficiencies are possible governmental interventions, all of which have different effects. Tailoring the regulation to the industry and the desired effect allows a flexible approach which preserves an industry's competitive edge. So, for example, a carbon tax on the national steel industry designed to reduce fossil fuel use could be coupled with a carbon tariff on steel imported from non-compliant foreign manufacturers, thus leveling the playing field. The degree of government intervention should also be considered. It has been suggested that direct funding is more appropriate to the early development phase of a new energy technology whereas indirect measures such as tax breaks are more suitable for encouraging industries to introduce new technology into the market place (Deutch and Lester 2004). From an economic perspective it seems reasonable to reduce all forms of regulation (aside from safety issues) for new energy technologies as they reach the point at which they can successfully compete in the market place. A slow phasing in of mandated efficiencies over long time periods allows industries to make long range plans, an important factor considering the 40–50 year lifetime of a typical power plant.

Energy use and gross domestic product (GDP) both for the US and the world have steadily increased over the past 50 years, which would seem to indicate energy is required for economic growth. Figure 10.4, however, shows that the ratio of

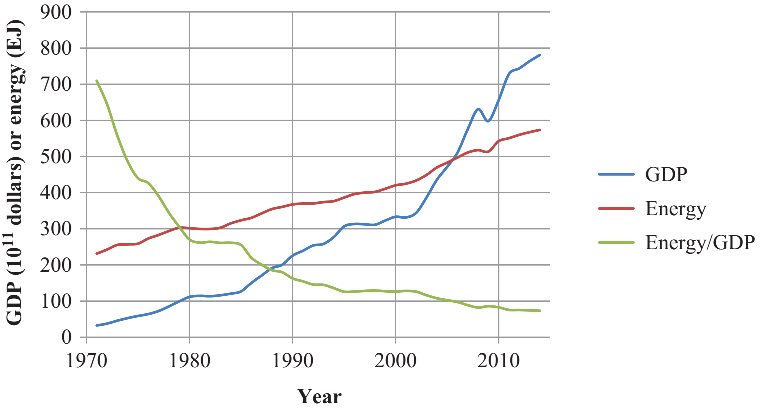


Figure 10.4. World energy use (in EJ) versus GDP (in constant year 2000 US dollars times 10^{11}). The ratio of energy use to GDP (times 100) is also shown. Data are from the World Bank (WB 2016) and IEA (2016).

energy to GDP, the *energy intensity* of the global economy, has been declining for the past 50 years, which indicates that the energy needed for a given increase of GDP is not a constant figure.

Per capita energy use is still increasing in developing countries but has been more or less constant in developed countries for the past 30 years, which is in part due to technological advancements. As an example of past technological improvement, refrigerators are 75% more efficient than they were in 1970 but are 20% larger and cost 60% less in today's dollars[14]. This is expected to continue as new technologies mature.

The automobile industry strongly lobbied against the original CAFE standards shown in figure 8.1, arguing that requiring additional energy economy would increase the cost of automobiles and make them less competitive in the global market. The success of Asian and European car manufacturers with cars with higher fuel efficiencies (mainly as a reaction to high gasoline taxes in Asia and Europe) in the American market is a counter-example, however. In spite of the additional manufacturing costs involved in making a car with higher fuel economy, these cars are very competitive in US and other markets. It is estimated that an additional 50% increase in fuel economy is possible by applying existing technology.

Economic barriers to increased efficiencies in construction include the fact that buildings are not built by the same entities which will use the building, so there is an incentive to cut construction cost at the expense of incorporating energy saving measures. Often the occupant of the building does not notice the additional energy expense because it is relatively low compared to mortgage or rent expenses, even though in the long term the energy needs for operating the building over its lifetime are generally a factor of ten higher than the energy needed to construct the building

---

[14] Unfortunately the gains from energy efficient refrigerators has largely been negated in the US by increased use of personal electronics in the past 15 years (Glicksman 2008).

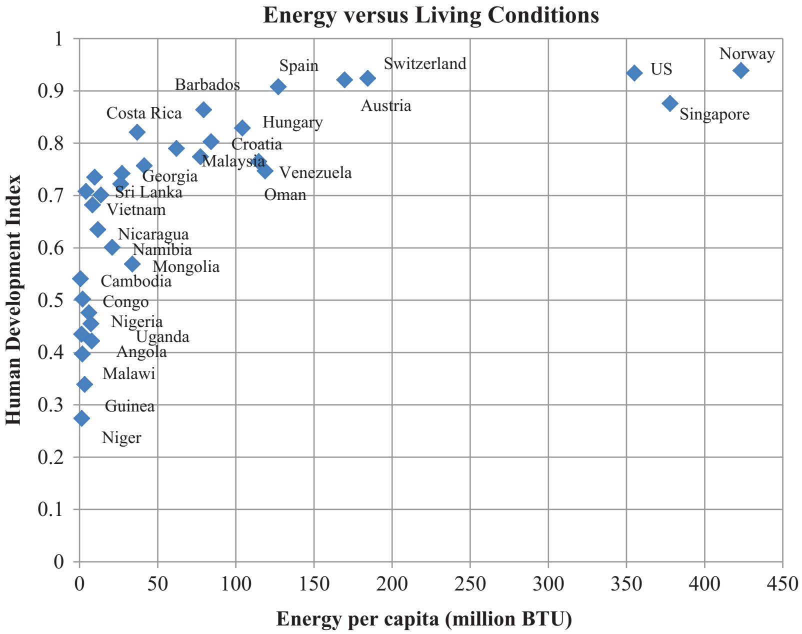


Figure 10.5. HDI versus per capita energy consumption. From Forinash (2010). Copyright 2010, Kyle Forinash. Reproduced by permission of Island Press, Washington DC.

(Glicksman 2008). Tax incentives and mandated efficiencies are ways to circumnavigate this problem.

Another measure of energy conservation and its economic cost is to look at standard of living versus energy use, shown in figure 10.5. The United Nations Human Development Index (HDI) combines economic well-being (GDP per capita) with access to education (measured as adult literacy and percentage school enrollment), and longevity which is related to health (UN 2015). The inclusion of these additional factors is an attempt to circumvent the problem that wealth does not necessarily correlate well with health or self-reported happiness. Several other measures for measuring well-being have been proposed, such as gross national product, net national welfare, gross domestic happiness, index of sustainable economic welfare, genuine progress indicator, sustainable net benefit index and green GDP, all of which take into account various social and economic factors of happiness, well-being, progress and growth.

In figure 10.5 we see that a per capita energy use of less than 100 million BTU is sufficient to raise the HDI to a level above 85%. It is also obvious from the graph that some countries, for example Norway and the US, have approximately the same HDI as Spain and other countries which use one third the energy per person. One could argue that the US requires more energy because the transportation distances are longer than in many other parts of the world, however, an almost identical graph can be made by only looking at *electrical* use per capita (Hafemister 2007). This clearly shows it is *not* the longer driving distances in the US that cause the difference in energy per capita, but rather other aspects involving lifestyle. Clearly the quality of life as measured by the HDI does not depend strongly on energy use per capita after some minimum amount which appears to be much less than that used in the US.

Projects

1. Write a report comparing various causes of death as they have changed over time and for different countries in the world. Point out any unusual or interesting facts about death rates around the globe. In particular, find data that show the increase in death rate due to smoking for males started a decade or more before females. Explain this trend based on social values of the time.
2. Write a report of how the economic risk of climate change is calculated (see Nordhaus 2013 as a start).
3. Write a summary of each of the major nuclear disasters which have occurred. Include estimates (from reliable sources) of injuries, deaths and LOLE. Compare these deaths to the fatality rate of automobiles and from coal fired plants.
4. Go to the NASA (2016) Near Earth Object Program http://neo.jpl.nasa.gov/ website and report on the chances of a near Earth object colliding with the Earth.
5. Starting with Roth and Ambs (2004) write a report on energy industry externalities which have been identified. Include an explanation of how the cost of these externalities are determined.
6. Download the DICE model (www.econ.yale.edu/~nordhaus/homepage/DICEmodels09302016.htm), run the model and report on various factors that strongly influence the necessary investment to limit an increase in global temperature to a few degrees.
7. Duplicate figure 10.4 but for specific countries and compare the results. (Hint: Check the World Bank data, http://data.worldbank.org/, or the CIA Factbook www.cia.gov/library/publications/the-world-factbook/.)
8. Write a report on the various different ways of measuring well-being mentioned in the text. For one or more of these factors, make a graph versus energy use per capita and compare with figure 10.5.

References

CDC (Center for Disease Control) 2015 Motor vehicle crash deaths www.cdc.gov/vitalsigns/motor-vehicle-safety/index.html (Accessed: 15 December 2016)

Chapman C R and Morrison D 1994 Impacts on the Earth by asteroids and comets: assessing the hazard *Nature* **367** 33

Cohen B L 1987 The nuclear reactor accident at Chernobyl, USSR *Am. J. Phys* **55** 1076

Crets J, Derkach A, Nyquist S, Ostrowski K and Stephanson J 2007 Reducing US greenhouse gas emissions: how much at what cost?' *McKinsey & Company Report* www.mckinsey.com/

business-functions/sustainability-and-resource-productivity/our-insights/reducing-us-greenhouse-gas-emissions (Accessed: 15 December 2016)

Dasgupta P S and Ehrlich P R 2013 Pervasive externalities at the population, consumption, and environment nexus *Science* **340** 324

Deutch J M and Lester R K 2004 *Making Technology Work* (Cambridge: Cambridge University Press)

EEA (European Environmental Agency) 2007 EN35 External costs of electricity production www.eea.europa.eu/data-and-maps/indicators/en35-external-costs-of-electricity-production-1 (Accessed: 15 December 2016)

EIA (Energy Information Administration) 2015 Direct Federal Financial Interventions and Subsidies in Energy in Fiscal Year 2013 www.eia.gov/analysis/requests/subsidy/ (Accessed: 15 December 2016)

Energy Institute (University of Texas, Austin) 2017 The full cost of electricity http://energy.utexas.edu/the-full-cost-of-electricity-fce/ (Accessed: 2 January 2017)

EPA (Environmental Protection Agency) 2015 The social cost of carbon www.epa.gov/climatechange/social-cost-carbon (Accessed: 16 December 2016)

Forinash K 2010 *Foundations of Environmental Physics* (Washington DC: Island Press)

IEA (International Energy Agency) 2016 Statistics www.iea.org/statistics/ (Accessed: 16 December 2016)

Glicksman L R 2008 Energy efficiency in the built environment *Phys. Today* **61** 35

Hafemister D 2007 *Physics of Societal Issues* (New York: Springer)

Harte J 1988 *Consider a Spherical Cow* (South Orange, NJ: University Science)

Kammen D M and Hassenzal D M 1999 *Should We Risk It? Exploring Environmental, Health, and Technological Problem Solving* (Princeton, NJ: Princeton University Press)

NASA 2016 Near Earth Object Program http://neo.jpl.nasa.gov/ (Accessed: 15 December 2016)

National Academy of Science 2010 *Electricity From Renewable Resources: Status, Prospects, and Impediments* www.nap.edu/catalog/12619/electricity-from-renewable-resources-status-prospects-and-impediments (Accessed: 15 December 2016)

Nordhaus W D 2013 *The Climate Casino* (New Haven, CA: Yale University Press)

NRC (National Research Council) 2010 Hidden costs of energy: unpriced consequences of energy production and use *Committee on Health, Environmental, and Other External Costs and Benefits of Energy Production and Consumption* www.nap.edu/download/12794 (Accessed: 16 December 2016)

NREL (National Renewable Energy Laboratory) 2016 Life Cycle Assessment Harmonization www.nrel.gov/analysis/sustain_lcah.html (Accessed: 15 December 2016)

NSC (US National Safety Council) 2016 Crashes involving cell phones *White paper* www.nsc.org/DistractedDrivingDocuments/NSC-Under-Reporting-White-Paper.pdf (Accessed: 6 January 2017)

Randolph J and Masters G M 2008 *Energy for Sustainability* (Washington, DC: Island)

Roth I F and Ambs L L 2004 Incorporating externalities into a full cost approach to electric power generation life-cycle costing *Energy* **29** 2125

Stern N 2007 *The Stern Review on the Economics of Climate Change* (Cambridge: Cambridge University Press)

Tester J W, Drake E M, Driscoll M J, Golay M W and Peters W A 2005 *Sustainable Energy: Choosing Among Options* (Cambridge, MA: MIT Press)

United Nations 2015 Human Development Data http://hdr.undp.org/en/data (Accessed: 16 December 2016)

WB (The World Bank) 2016 http://data.worldbank.org/indicator/NY.GDP.MKTP.CD (Accessed: 16 December 2016)

Wilson R and Crouch E A C 2001 *Risk-Benefit Analysis* (New Haven, CA: Harvard University Press)

Wilson R 2002 Resource letter: RA-1: risk analysis *Am. J. Phys.* **70** 475