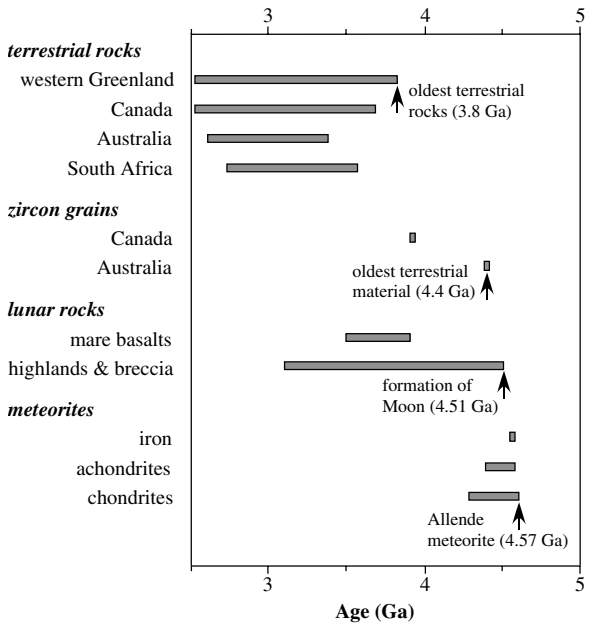

Fig. 4.12 Radiometric age ranges of the oldest terrestrial and lunar rocks and meteorites (compiled from Dalrymple, 1991 and Halliday, 2001).

source areas of the rocks. The dark areas of the Moon's surface, the so-called lunar seas or *maria*, were formed when enormous outpourings of basaltic lava filled up low-lying areas – such as the craters formed by large meteoritic impacts – and so created flat surfaces. The lunar volcanism may have persisted until 1 Ga ago. The light areas on the Moon's surface are rough, extensively cratered highlands that reach elevations of 3000–4000 m above the maria. They represent the top part of the lunar crust and are the oldest regions of the Moon. Frequent collisions with large asteroids early in the Moon's history produced numerous craters and pulverized the lunar crust, leaving impact breccia, rock fragments and a dust layer a few meters thick, called the lunar *regolith*. Age dates from the highland rocks range from about 3.5–4.5 Ga, but the oldest age reported for a lunar rock is 4.51 ± 0.07 Ga, obtained by the Rb–Sr method.

In contrast to the ages of meteorites and the Moon, no rock of comparable age has been preserved on the Earth. Among the oldest terrestrial rocks are the Isua metasediments from western Greenland. They have been extensively studied and yield an age of 3.77 Ga that is consistent with different decay schemes. Even older are the Acasta gneisses in the northwestern part of the Canadian shield, which gave a U–Pb age of 3.96 Ga. Terrestrial rocks from the oldest Precambrian shield areas in Australia, South Africa, South America and Antarctica give *maximum* ages of 3.4–3.8 Ga.

Older ages are obtained from individual grains of zircon ($ZrSiO_4$), which forms as a trace mineral in granitic magmas. Zircon is a very durable mineral and can survive erosion that destroys the original rock. The zircon grains then become incorporated in sedimentary rocks. Zircon grains from sedimentary rocks in Australia have yielded U–Pb ages of 4.1–4.2 Ga; a fragment of a zircon grain, only 200 μm in diameter, has been dated by the U–Pb method at 4.404 ± 0.008 Ga, making it the oldest dated terrestrial solid. It is also possible to carry out oxygen-isotope analyses on these ancient zircon grains. Although fragmentary, the results from zircon grains are important for understanding conditions on the primeval Earth. The scanty evidence suggests a possible history for the development of the hot early Earth, as follows.

The Sun and its accretionary disk, from which the solar system eventually evolved, formed 4.57 Ga ago. The accretion of the Earth and formation of its iron core lasted for about the first 100 Ma of Earth's early history. The accretion of Mars, which has only about one-ninth the mass of Earth, would have been completed in the first 30 Ma. Later in Earth's accretion, after perhaps 60 Ma, the Moon was formed as a result of the collision – called the "Giant Impact" – of a Mars-sized planetesimal with the proto-Earth, which was still much smaller than its present size. In the interval up to 4 Ga ago – called the *Hadean* (for hell-like) era – the early Earth was hot and covered by magma oceans; any crust would melt or be destroyed in the intense meteoritic bombardment. Water, if present, would be vaporized. After about 150 Ma the proto-Earth might have cooled sufficiently so that an early granitic crust, and possibly liquid water, could form. Only a few zircon grains that survived this period remain as possible witnesses. Repeated bombardments may have destroyed the crust and vaporized the water repeatedly in the following 400 Ma. Around 4 Ga ago the oldest surviving continental crust was formed.

This scenario is speculative and it is not unique. Some oxygen-isotope studies on zircon grains have been interpreted as indicating a cool early Earth. High $\delta^{18}O$ values (see Box 5.2) measured in these zircons imply surface temperatures low enough for liquid water; uniform conditions throughout the Archean era (4.4–2.6 Ga) were deduced. This model of early Earth suggests that meteoritic impacts may have been less intense than usually hypothesized. The model of a hot early Earth is more popularly accepted but the true history is not yet known. The events that happened in the first few hundred million years of its history are among Earth's best-kept secrets.

4.2 THE EARTH'S HEAT

4.2.1 introduction

The radiant energy from the Sun, in conjunction with gravitational energy, determines almost all natural processes that occur at or above the Earth's surface. The hot incandescent Sun emits radiation in a very wide range of wavelengths. The radiation incident on the Earth is largely reflected into space, part enters the atmosphere and is reflected by the clouds or is absorbed and re-radiated into space. A very small part reaches the surface, where it is also partly reflected, especially from the water

surfaces that cover three-quarters of the globe. Some is absorbed (e.g., by vegetation) and serves as the source of power for various natural cycles. A small fraction is used to heat up the Earth's surface, but it only penetrates a short distance, some tens of centimeters in the case of the daily cycle and a few tens of meters for the annual changes. As a result, solar energy has negligible influence on internal terrestrial processes. Systems as diverse as the generation of the geomagnetic field and the motion of global lithospheric plates are ultimately powered by the Earth's internal heat.

The Earth is constantly losing heat from its interior. Although diminutive compared to solar energy, the loss of internal heat is many times larger than the energy lost by other means, such as the change in Earth's rotation and the energy released in earthquakes (Table 4.3). Tidal friction slows down the Earth's rotation, and the change can be monitored accurately with modern technology such as very long baseline interferometry (VLBI) and the satellite-based geodetic positioning system (GPS). The associated loss of rotational energy can be computed accurately. The elastic energy released in an earthquake can be estimated reliably, and it is known that most of the energy is released in a few large shocks. However, the annual number of large earthquakes is very variable. The number with magnitude $M_s > 7$ varies between about 10 and 40 (see Fig. 3.49), giving estimates of the annual energy release from about 5×10^{17} J to 4×10^{19} J. The energies of tidal deceleration and earthquakes are small fractions of the geothermal flux, which is the most important form of energy originating in the body of the Earth.

The Earth's internal heat derives from several sources (Section 4.2.5). For the past 4 Ga or so the Earth's heat has been obtained from two main sources. One is the cooling of the Earth since its early history, when internal temperatures were much higher than they now are. The other is the heat produced by the decay of long-lived radioactive isotopes. This is the main source of the Earth's internal heat, which, in turn, powers all geodynamic processes.

Table 4.3 *Estimates of notable contributions to the Earth's annual energy budget*

Energy source	Annual energy [J]	Normalized [geothermal flux = 1]
Reflection and re-radiation of solar energy	5.4×10^{24}	≈ 4000
Geothermal flux from Earth's interior	1.4×10^{21}	1
Rotational deceleration by tidal friction	$\approx 10^{20}$	≈ 0.1
Elastic energy in earthquakes	$\approx 10^{19}$	≈ 0.01

4.2.2 Thermodynamic principles

In order to describe thermal energy it is necessary to define clearly some important thermodynamic parameters. The concepts of temperature and heat are easily – and frequently – confused. Temperature – one of the seven fundamental standard parameters of physics – is a quantitative measure of the degree of hotness or coldness of an object relative to some standard. Heat is a form of energy which an object possesses by virtue of its temperature. The difference between temperature and heat is illustrated by a simple example. Imagine a container in which the molecules of a gas move around at a certain speed. Each molecule has a kinetic energy proportional to the square of its velocity. There may be differences from one molecule to the next but it is possible to determine the mean kinetic energy of a molecule. This quantity is proportional to the temperature of the gas. If we add up the kinetic energies of all molecules in the container we obtain the amount of heat it contains. If heat is added to the container from an external source, the gas molecules speed up, their mean kinetic energy increases and the temperature of the gas rises.

The change of temperature of a gas is accompanied by changes of pressure and volume. If a solid or liquid is heated, the pressure remains constant but the volume increases. Thermal expansion of a suitable solid or liquid forms the principle of the thermometer for measuring temperature. Although Galileo reputedly invented an early and inaccurate "thermoscope," the first accurate thermometers – and corresponding temperature scales – were developed in the early eighteenth century by Gabriel Fahrenheit (1686–1736), Ferchaut de Réaumur (1683–1757) and Anders Celsius (1701–1744). Their instruments utilized the thermal expansion of liquids and were calibrated at fixed points such as the melting point of ice and the boiling point of water. The Celsius scale is the most commonly used for general purposes, and it is closely related to the scientific temperature scale.

Temperature apparently has no upper limit. For example, the temperature of the surface of the Sun is less than 10,000 K but the temperature at its center is around 10,000,000 K and temperatures greater than 100,000,000 K have been achieved in physics experiments. But as heat is removed from an object it becomes more and more difficult to lower its temperature further. The limiting low temperature is often called "absolute zero" and is taken as the zero of the *Kelvin temperature scale*, named in honor of Lord Kelvin. Its divisions are the same as the Celsius scale and the temperature unit is called a *kelvin*. The scale is defined so that the triple point of water – where the solid, liquid and gaseous phases of water can coexist in equilibrium – is equal to 273.16 kelvins, written 273.16 K.

Heat was imagined by early investigators to be exchanged between bodies by the flow of a mystic fluid, called *caloric*. However, in the mid nineteenth century James Joule, an English brewer, demonstrated in a series of careful experiments that mechanical energy could be converted into heat. In his famous experiment, falling

weights drove a paddle wheel in a container of water, raising its temperature. The increase was tiny, less than 0.3 K, yet Joule was able to compute the amount of energy needed to raise the temperature by 1 K. His estimate of this energy – called the *mechanical equivalent of heat* – was within 5% of our modern value. The unit of energy is called the *joule* in recognition of his pioneering efforts. Originally, however, the unit of heat energy was defined as the amount needed to raise the temperature of one gram of water from 14.5 °C to 15.5 °C. This unit, the *calorie* (cal), is equivalent to 4.1868 J.

In physics and engineering it is often important to know the change of heat energy in a unit of time, known as the power. The unit of power is the *watt*, named after James Watt, the Scottish engineer who played an important role in harnessing thermal energy as a source of mechanical power. In geothermal problems we are usually concerned with the loss of heat from the Earth per unit area of its surface. This quantity is called the *heat flux* (or more commonly *heat flow*); it is the amount of heat that flows per second across a square meter of surface. The mean heat flow from the Earth is very small and is measured in units of milliwatt per square meter (mW m^{-2}).

The addition of a quantity of heat ΔQ raises the temperature by an amount ΔT, which is proportional to ΔQ. The larger the mass m of the body, the smaller is the temperature change, and a given amount of heat produces different temperature changes in different materials. The amount of heat needed to raise the temperature of 1 kg of a material by 1 K is called its *specific heat*, denoted c_p for a process that occurs at constant pressure (and c_v when it happens at constant volume). These observations are summarized in the equation

$$\Delta Q = c_p m \Delta T \tag{4.25}$$

The added heat causes a fractional change of volume that is proportional to the temperature change but which differs from one material to another. The material property is called the *volume coefficient of expansion* α, and is defined by the equation

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_p \tag{4.26}$$

When thermal energy is added to a system, part is used to increase the internal energy of the system – i.e., the kinetic energy of the molecules – and part is expended as work, for example, by changing the volume. If the change in total energy ΔQ occurs at constant temperature T, we can define a new thermodynamic parameter, the *entropy* S, which changes by an amount ΔS equal to $\Delta Q/T$. Thus we can write

$$\Delta Q = T\Delta S = \Delta U + \Delta W \tag{4.27}$$

where ΔU is the change of internal energy and ΔW is the work done externally. A thermodynamic process in which heat cannot enter or leave the system is said to be *adiabatic*. The entropy of an adiabatic reaction remains constant: $\Delta S = 0$. This is the case when a process occurs so rapidly that there is no time for heat transfer. An example is the passage of a seismic wave in which the compressions and rarefactions occur too rapidly for heat to be exchanged. The adiabatic temperature gradient in the Earth serves as an important reference for estimates of the actual temperature gradient and for determining how heat is transferred.

4.2.3 Temperature inside the Earth

In contrast to the radial distributions of density, seismic velocity and elastic parameters, which are known with a good measure of reliability, our knowledge of the temperature inside the Earth is still imprecise. The temperature can only be measured in the immediate vicinity of the Earth's surface, in boreholes and deep mines. As early as 1530 Georgius Agricola (the latinized name of Georg Bauer, a German physician and pioneer in mineralogy and mining) noted that conditions were warmer in deep mines. In fact, near-surface temperatures increase rapidly with depth by roughly 30 K km^{-1}. At this rate, linear extrapolation would give a temperature around 200,000 K at the center of the Earth. This is greater than the temperature of the surface of the Sun and is unrealistically high.

The conditions of high temperature and pressure in the deep interior can be inferred from experiments, and the adiabatic and melting-point temperatures can be computed with reasonable assumptions. Nevertheless, the temperature–depth profile is poorly known and conjectured temperatures have ranged widely. Limits are placed on the actual temperature by the known physical state of the Earth's interior deduced from seismology. The temperature in the solid inner core must be lower than the melting point, while the temperature of the molten outer core is above the melting point. Similarly the temperature in the solid mantle and crust are below the melting point; the asthenosphere has low rigidity because its temperature comes close to the solidus ("softening point"). The relationship of the actual temperature to the melting point determines how different parts of the Earth's interior behave rheologically (see Section 2.8).

The experimental approach to estimating the variation of temperature with depth combines knowledge obtained from seismology with laboratory results. The travel-times of seismic body waves show that changes in mineral structure (phase transitions) occur at certain depths (see Section 3.7.5). Important examples are the olivine–spinel transition at 400 km depth and the spinel–perovskite transition at 670 km depth in the upper mantle. The conditions of temperature and pressure (and hence depth) at which these phase transitions take place can be observed in laboratory experiments, so that the temperatures at the transition depths in the Earth can be determined. Similarly, the depth variation of the melting points of mantle rocks and the iron–nickel core can be inferred from laboratory observations at high pressure and temperature. Seismic

velocities in the Earth are now so well known that deviations from normal velocities can be determined by seismic tomography (see Section 3.7.6) and interpreted in terms of temperature anomalies.

4.2.3.1 The adiabatic temperature gradient

An alternative way of estimating temperature inside the Earth is by using physical equations in which the parameters are known from other sources. In the late nineteenth century James Clerk Maxwell expressed the laws of thermodynamics in four simple equations involving entropy (S), pressure (p), temperature (T) and volume (V). One of these equations is

$$\left(\frac{\partial T}{\partial p}\right)_{\mathrm{S}} = \left(\frac{\partial V}{\partial S}\right)_{\mathrm{p}} \tag{4.28}$$

The left side is the adiabatic change in temperature with pressure, from which we obtain the adiabatic change in temperature with depth by substituting $\mathrm{d}p = \rho g\ \mathrm{d}z$, as in Section 3.7.4. Substituting from Eqs. (4.25) and (4.26) we get

$$\left(\frac{\partial T}{\partial p}\right)_{\mathrm{S}} = T\frac{\alpha V \Delta T}{c_{\mathrm{p}} m \Delta T} = T\frac{\alpha}{\rho c_{\mathrm{p}}} \tag{4.29}$$

from which we obtain

$$\left(\frac{\partial T}{\partial z}\right)_{\text{adiabatic}} = T\frac{\alpha g}{c_{\mathrm{p}}} \tag{4.30}$$

The dependence of density and gravity on depth z are known from seismic travel-times, and the profiles of α and c_{p} can be estimated from laboratory observations (Fig. 4.13). For example, in the lower mantle at a depth of 1500 km, $g = 9.9\,\mathrm{ms}^{-2}$, $c_{\mathrm{p}} = 1200\,\mathrm{J\,kg^{-1}\,K^{-1}}$, $\alpha = 14 \times 10^{-6}\ \mathrm{K^{-1}}$, and $T = 2400\,\mathrm{K}$. This gives an adiabatic temperature gradient of about $0.3\,\mathrm{K\,km^{-1}}$. In the outer core at about 3300 km depth the corresponding values are: $g = 10.1\,\mathrm{m\ s^{-2}}$, $c_{\mathrm{p}} = 700\,\mathrm{J\ kg^{-1}\ K^{-1}}$, $\alpha = 14 \times 10^{-6}\ \mathrm{K^{-1}}$, and $T = 4000\,\mathrm{K}$ and the adiabatic temperature gradient is about $0.8\,\mathrm{K\ km^{-1}}$.

Approximate estimates of adiabatic temperatures inside the Earth can also be obtained with the aid of the Grüneisen thermodynamic parameter, γ. This is a dimensionless parameter, defined as

$$\gamma = \frac{\alpha K_{\mathrm{s}}}{\rho c_{\mathrm{p}}} \tag{4.31}$$

where K_{s} is the adiabatic incompressibility or bulk modulus. It is defined in Section 3.2.4 and Eq. (3.17), which, by writing $\mathrm{d}p$ instead of p and $\mathrm{d}V/V$ for the dilatation θ, becomes

$$\mathrm{d}p = -K_{\mathrm{s}}\frac{\mathrm{d}V}{V} = K_{\mathrm{s}}\frac{\mathrm{d}\rho}{\rho} \tag{4.32}$$

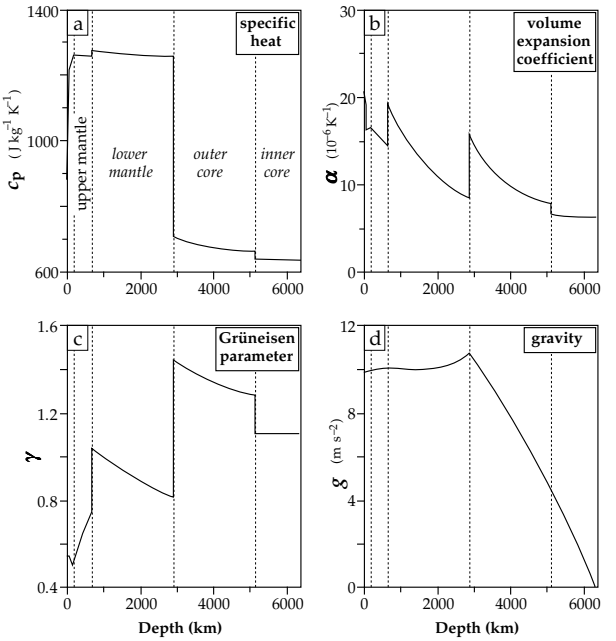

Fig. 4.13 Variations with depth in the Earth of (a) specific heat at constant pressure, (b) volume coefficient of thermal expansion, (c) Grüneisen parameter, and (d) gravity (based upon data from Stacey, 1992).

where ρ is the density. Substituting $K_{\mathrm{s}} = \rho\ \partial p/\partial\rho$ in Eq. (4.29) gives

$$\frac{\mathrm{d}T}{\mathrm{d}p} = T\frac{\gamma}{K_{\mathrm{s}}} = T\frac{\gamma\,\mathrm{d}\rho}{\rho\,\mathrm{d}p}$$

$$\frac{\mathrm{d}T}{T} = \gamma\frac{\mathrm{d}\rho}{\rho} \tag{4.33}$$

$$T = T_0\left(\frac{\rho}{\rho_0}\right)^{\gamma} \tag{4.34}$$

With this equation, and knowing the temperature T_0 and density ρ_0 at a given depth, the adiabatic temperature can be computed from the density profile in a region where the Grüneisen parameter γ is known. Fortunately, γ is fairly constant within large regions of the Earth's interior (Fig. 4.13). Clearly, Eq. (4.34) cannot be applied across a boundary between these domains, where γ is discontinuous. If T_0 and ρ_0 are known at calibration points, the adiabatic temperature profile may be computed iteratively within a particular depth interval. A current estimate of the temperature profile in the Earth (Fig. 4.14) has steep gradients in the lithosphere, asthenosphere and in the D″ layer above the core–mantle boundary. It indicates a temperature near 3750 K at the core–mantle boundary and a central temperature of about 5100 K.

4.2.3.2 The melting point gradient

Another of Maxwell's thermodynamic equations is

$$\left(\frac{\partial S}{\partial p}\right)_{\mathrm{T}} = -\left(\frac{\partial V}{\partial T}\right)_{\mathrm{p}} \tag{4.35}$$

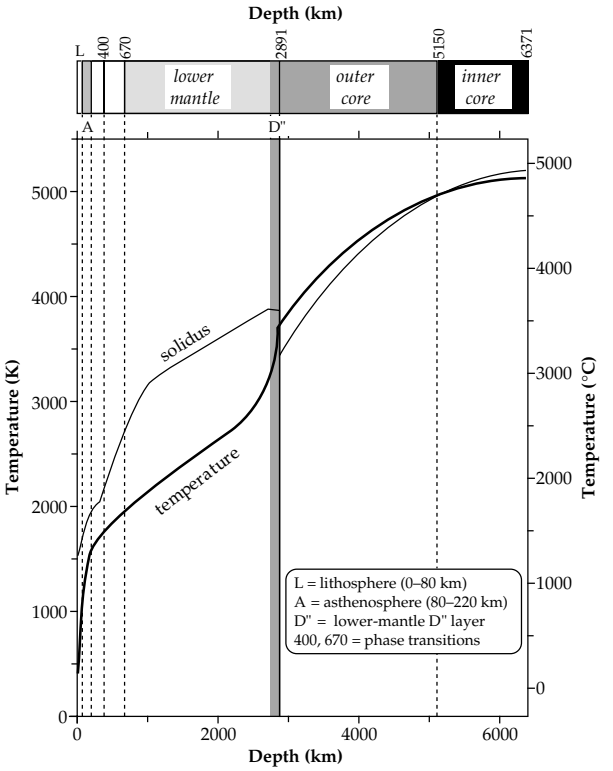

Fig. 4.14 Variations of estimated temperature and melting point with depth in the Earth (based upon data from Stacey, 1992).

This equation can be applied to the effect of pressure on the melting point of a substance (T_{mp}). The heat required to melt a unit mass of the substance is its latent heat of fusion (L), so the change in entropy on the left-hand side of the equation is equal to (mL/T_{mp}). The volume change is the difference between that of the solid phase (V_S) and that of the liquid phase (V_L), so Eq. (4.35) can be rewritten

$$\frac{dT_{mp}}{dp} = \frac{T_{mp}}{mL}(V_S - V_L) \tag{4.36}$$

This is known to physicists as the *Clausius–Clapeyron equation*. It describes the effect of pressure on the melting point, and it is of interest to us because we can easily convert it to give the variation of melting point with depth, assuming that the pressure is hydrostatic so that $dp = \rho g\, dz$ as previously. For a given mass m of the substance we can replace the volumes V_S and V_L with the corresponding densities ρ_S and ρ_L of the solid and liquid phases, respectively, so that

$$\frac{1}{T_{mp}}\frac{dT_{mp}}{dz} = \frac{g}{L}\left(\frac{\rho_S}{\rho_L} - 1\right) \tag{4.37}$$

Again, to obtain the depth distribution of the melting point the variations of density and gravity with depth in the Earth are needed. At outer core pressures the densities of the solid and liquid phases of iron are about 13,000 kg m^{-3} and 11,000 kg m^{-3}, respectively, and the latent heat of fusion of iron is about 7×10^6 J kg^{-1}, so the gradient of the melting point curve in the outer core is about 1 K km^{-1}, i.e., the melting point in the core increases more steeply with depth than the adiabatic temperature. The computations of the adiabatic and melting temperature curves depend on parameters (e.g., L, α, c_p) that are not known with a great degree of reliability in the Earth so the temperature profiles (Fig. 4.14) will undoubtedly change and become more secure as basic knowledge improves.

One factor that must still be evaluated is the role of phase transitions in the mantle. The D″ layer just above the core–mantle boundary evidently plays a crucial role in transferring heat from the core to the mantle. It constitutes a thermal boundary layer. Likewise the lithosphere forms a thermal boundary layer that conveys mantle heat to the Earth's surface. It appears unlikely that the phase transition at 400 km constitutes a thermal boundary layer but the phase transition at 670 km depth may do so. In the model used to derive the temperature profiles in Fig. 4.14 the phase transitions do not act as thermal boundary layers. Throughout most of the mantle the temperature gradient is assumed to equal the adiabatic gradient, but the mantle is bounded at top and bottom by thermal boundary layers (the lithosphere and D″-layer, respectively) in which the temperature gradient greatly exceeds the adiabatic gradient.

4.2.4 Heat transport in the Earth

Heat can be transported by three processes: conduction, convection and radiation. Conduction and convection require the presence of a material; radiation can pass through space or a vacuum. Conduction is the most significant process of heat transport in solid materials and thus it is very important in the crust and lithosphere. However, it is an inefficient form of heat transport, and when the molecules are free to move, as in a fluid or gas, the process of convection becomes more important. Although the mantle is solid from the standpoint of the rapid passage of seismic waves, the temperature is high enough for the mantle to act as a viscous fluid over long time intervals. Consequently, convection is a more important form of heat transfer than conduction in the mantle. Convection is also the most important form of heat transport in the fluid core, where related changes in the geomagnetic field show that the turnover of core fluid is rapid in geological terms. Radiation is the least important process of heat transport in the Earth. It is only significant in the hottest regions of the core and lower mantle. The absorption of radiant energy by matter increases its temperature and thereby the temperature gradient. Hence, thermal radiation can be taken into account as a modification of the ability of the material to transfer heat by conduction.

4.2.4.1 Conduction

Thermal conduction takes place by the transfer of kinetic energy between molecules or atoms. A true understanding

of the processes involved would require us to invoke quantum theory and the so-called "band theory of solids," but a general understanding is possible without resorting to such measures. The electrons in an atom that are most loosely bound – the valence electrons – are essentially free of the ionic cores and can move through a material, so transferring kinetic energy. Hence they are called *conduction electrons*. Because electrons are electrically charged, the net movement of conduction electrons also causes an electrical current. Not surprisingly materials that are good electrical conductors (e.g., silver, copper) also conduct heat well. In this atomic view the conduction electrons move at very high speeds ($\sim$1000 km s^{-1}) but in random directions so that there is no net energy transfer in any particular direction. In an electrical or temperature field the conduction electrons drift systematically down the slope of the field (i.e., in the direction of the electrical field or temperature gradient). The additional drift velocity is very small (about 0.1 mm s^{-1}) but it passes kinetic energy through the material. This form of conduction is possible in liquids, gases or solids.

An additional mechanism plays an important role in conduction in solids. The atoms in a solid occupy definite positions that form a lattice with a certain symmetry. The atoms are not stationary but vibrate at a frequency that is temperature dependent. The lattice vibrational energy is quantized, forming units called *phonons*. An increase in temperature at one end of a solid raises the lattice vibrational frequency there. Due to the coupling between atoms the increased vibration is eventually passed through the lattice as an increase in temperature.

The relative importance of electrons and phonons in conducting heat differs from one solid to another. In metals, which contain large numbers of conduction electrons, thermal transport is due largely to the electrons; the lattice conductivity is barely measurable. In an insulator or poor conductor, such as the minerals of the crust and mantle, there are few conduction electrons and thermal conductivity is largely determined by lattice vibrations (phonons).

The transport of heat by conduction in a solid is governed by a simple equation. Consider a solid bar of length L and cross-sectional area A with its ends maintained at temperatures T_1 and T_2, respectively (Fig. 4.15a). Assuming that heat flows only along the bar (i.e., there are no side losses) the net amount of heat (ΔQ) that passes in a given time from the hot end to the cold end depends directly on the temperature difference ($T_2 - T_1$), the area of cross-section (A) and the time of observation (Δt), and inversely on the length of the bar (L). These observations can be summarized in the equation

$$\Delta Q = kA\frac{T_2 - T_1}{L}\Delta t \tag{4.38}$$

The constant of proportionality, k, is the *thermal conductivity*, which is a property of the material of the bar. If the length of the bar is very small or the temperature

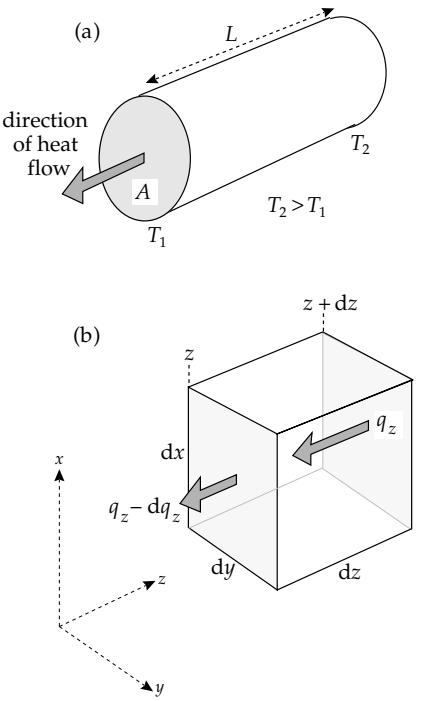

Fig. 4.15 (a) Conduction of heat Q through a bar of length L and cross-sectional area A, with its ends kept at temperatures T_1 and T_2 ($>T_1$). (b) Heat flux entering (q_z) and leaving ($q_z - dq_z$) a short bar of length dz.

change across it is uniform, the ratio $(T_2 - T_1)/L$ is the temperature gradient. We can modify the equation to describe the vertical flow of heat out of the Earth by substituting the vertical temperature gradient, (dT/dz), which is also called the *geothermal gradient*. Equation (4.38) can then be rearranged as follows

$$q_z = -\frac{1}{A}\frac{dQ}{dt} = -k\frac{dT}{dz} \tag{4.39}$$

In this equation q_z is the *heat flux*, defined as the flow of heat per unit area per second. The negative sign is needed to account for the direction of the heat flow; if temperature increases in the downward direction of the z-axis, the flow of heat from high to low temperature is upward. The mean value of the undisturbed geothermal gradient near to the Earth's surface is about 30 °C km^{-1}, with low values of around 10 °C km^{-1} in ancient crust and high values of around 50 °C km^{-1} in young active zones.

The change of temperature within a body is described by the heat conduction equation, which is solved in Section 4.2.6 for special situations that are of interest for the transfer of thermal energy in the Earth. Conduction is a slow, and less effective means of heat transport than convection. It is important in the rigid crust and lithosphere, where convection cannot take place. However, it cannot be neglected in the fluid core, which is metallic and therefore a good conductor. A significant part of the core's heat is conducted out of the core along the adiabatic temperature gradient. The remainder, in excess of the conductive heat flow, is transported by convection currents.

4.2.4.2 Convection

Suppose that a small parcel of material at a certain depth in the Earth is in thermal equilibrium with its surroundings. If the parcel is displaced vertically upward without gaining or losing heat, it experiences a drop in pressure accompanied by a corresponding loss in temperature. If the new temperature of the parcel is the same as that of its surroundings at the new depth, the conditions at each depth are in adiabatic equilibrium. The variation of temperature with depth then defines the adiabatic temperature curve.

Now suppose that the real temperature increases with depth more rapidly than the adiabatic temperature gradient. The temperature loss of the upwardly displaced parcel is due to the change in pressure, which will be the same as in the previous case. But the real temperature has dropped by a larger amount, so the parcel is now hotter and therefore less dense than its surroundings. Its buoyancy causes it to continue to rise until it reaches a level where it is in equilibrium or can rise no further. Meanwhile, the volume vacated by the displaced parcel is occupied by adjacent material. Conversely, if a parcel of material is displaced downward, it experiences adiabatic increases in pressure and temperature. The temperature increase is less than required by the real temperature gradient, so the parcel remains cooler than its surroundings and sinks further. A pattern of cyclical behavior arises in which material is heated up and rises, while cooler material sinks to take its place, and is in turn heated up and rises, and so on. The process is called *thermal convection* and the physical transportation of material and heat is called a *convection current*.

The difference between the real and adiabatic temperature gradients is the *superadiabatic* temperature gradient, θ. For thermal convection to take place in a fluid, θ must be positive. Suppose that the temperature at a certain depth exceeds the adiabatic temperature by an amount ΔT. The temperature excess causes a volume V of fluid to expand by an amount proportional to the volume coefficient of expansion, α; this causes a mass deficiency of ($V\rho\alpha\ \Delta T$). Archimede's principle applies, so the hot volume V experiences a buoyancy force given by:

$$F_B = V\rho g\alpha\Delta T \tag{4.40}$$

Two effects inhibit the hot volume from rising. First, some of the heat that would contribute to the buoyancy is removed by thermal conduction; the efficacy of this process is expressed by the *thermal diffusivity* κ of the material, which depends on its density ρ, thermal conductivity k and specific heat c_p (see Section 4.2.6). Second, as soon as the overheated volume of fluid begins to rise, it experiences a resisting drag due to the *viscosity* η of the fluid. The effects combine to produce a force, proportional to $\kappa\eta$, which opposes convection. If the volume V involved in the convection has a typical dimension D, so that $V \sim D^3$, we can define a dimensionless number Ra, the Rayleigh number, which is proportional to the ratio of the buoyancy force to the diffusive–viscous force:

$$\mathrm{Ra} = \frac{g\rho\alpha\Delta T}{\kappa\eta}D^3 \tag{4.41}$$

Initially, heat passes through the material by conduction, but the diffusion takes some time. If the heat flux is large enough, it cannot diffuse entirely. The temperature rises above the adiabatic and buoyancy forces develop. For convection to occur, the buoyancy forces must dominate the resisting forces. This does not happen until the Rayleigh number exceeds a critical value, which is determined additionally by the boundary conditions and the geometry of the convection. For example, the condition for the onset of convection in a thin horizontal fluid layer, heated from below and with the top and bottom surfaces free from stress, was shown by Lord Rayleigh in 1916 to depend on the value of Ra given by

$$\mathrm{Ra} = \frac{g\alpha\theta}{\kappa\nu}D^4 \tag{4.42}$$

Here D is the layer thickness, θ is the superadiabatic temperature gradient and ν (equal to $\eta\rho$) is the *kinematic viscosity*. Convection begins in the flat layer if Ra is greater than $27\pi^4/4 = 658$. In cases with different boundary conditions, or for convection to occur in a spherical shell, the critical Rayleigh number is higher. However, convection generally originates if Ra is of the order of 10^3 and when Ra reaches around 10^5 heat transport is almost entirely by convection with little being transferred by diffusion.

For convection to occur, the real temperature gradient must exceed the adiabatic gradient. However, the loss of heat by convection reduces the difference between the gradients. Accordingly, the adiabatic gradient evolves as a convecting fluid cools. An important effect of convection is to keep the temperature gradient close to the adiabatic gradient. This condition is realized in the Earth's fluid core, where convection is the major mechanism of heat transport. Thermal convection is augmented by *compositional convection* related to the solidification of the inner core. The core fluid is made up of iron, nickel and lower-density elements, e.g., sulfur. Solidification of the inner core separates the dense iron from the lower-density elements at the inner core boundary. Being less dense than the core fluid, the residual materials experience an upward buoyancy force, resulting in a cycle of compositionally driven convection. Thermal and compositional convection in the Earth's core each act as a source of the energy needed to drive the geomagnetic field, with compositional convection the more important type.

Convection is the most important process of thermal transport in the fluid core, but it is also important in the mantle. The material of the Earth's mantle is rigid to the short-lived passage of seismic waves but is believed to yield slowly over long periods of time (Section 2.8.6). Although the mantle viscosity is high, the time-scale of geological processes is so long that long-term flow can

take place. The flow patterns are dominated by thermal convection and are influenced by the presence of thermal boundary layers, which the flowing material cannot cross. However, convection is a more effective mechanism than conduction and it is thought to be the dominant process of heat transfer in the mantle (Section 4.2.9).

A further process of heat transfer that involves bodily transport of matter is *advection*. This can be regarded as a form of forced convection. Instead of being conveyed by thermally produced buoyancy, advected heat is transported in a medium that is itself driven by other forces. For example, in a thermal spring the flow of water is due to hydraulic forces and not to density differences in the hot water. Similarly, volcanic eruptions transport advected heat along with the lava flow, but this is propelled by pressure differences rather than by buoyancy.

4.2.4.3 Radiation

Atoms can exist in many distinct energy states. The most stable is the ground state, in which the energy is lowest. When an atom changes from an excited state to a lower-energy state, it is said to undergo a transition. Energy corresponding to the difference in energy between the states is emitted as an electromagnetic wave, which we call radiation. Quantum physics teaches that the radiant energy emitted consists of a discrete number of fundamental units, called *quanta*. The particular wavelength of the electromagnetic radiation associated with a transition is proportional to the energy difference between the two states. If several different transitions are taking place simultaneously, the body emits a spectrum of wavelengths. Radio signals, heat, light, and x-rays are examples of electromagnetic radiation that have different wavelengths. The electromagnetic wave consists of fluctuating electric and magnetic fields, which need no medium for their passage. For this reason, radiation can travel through space or a vacuum. In materials it may be scattered or absorbed, depending on its wavelength. Heat radiation corresponds to the *infrared* part of the electromagnetic spectrum with wavelengths just longer than those of visible light.

The radiation of a commonplace hot object depends on factors that are difficult to assess. Classical physics fails to explain adequately the absorption and emission of radiation. To provide an explanation physicists introduced the concept of a *black body* as a perfect absorber and emitter of radiation. At any temperature it emits a continuous spectrum of radiation; the frequency content of the spectrum does not depend on the material composition of the body but only on its temperature. An ideal black body does not exist in practice, but it can be approximated by a hollow container that has a small hole in its wall. When the container is heated, the radiation escaping through the hole – so-called cavity radiation – is effectively black-body radiation. In 1879 Josef Stefan pointed out that the loss of heat by radiation from a hot object is proportional to the fourth power of the absolute temperature. If R represents the radiant energy per second emitted per unit area of the surface of the body at temperature T, then

$$R = \sigma T^4 \tag{4.43}$$

where σ, known as Stefan's constant or the Stefan–Boltzmann constant, has the value 5.6704×10^{-8} W m^{-2} K^{-4}.

In 1900, Max Planck, professor of physics at the university of Berlin, proposed that an oscillator could only have discrete amounts of energy. This was the birth of quantum theory. The energy of an oscillator of frequency ν is equal to the product $h\nu$, where the universal constant h (known as Planck's constant) has the value 6.626×10^{-34} J s. The application of quantum principles to black-body radiation provides a satisfactory explanation of Stefan's law and allows Stefan's constant to be expressed in terms of other fundamental physical constants.

Radiation is reflected and refracted in a transparent medium wherever the refractive index n changes; energy is transferred to the medium in each of these interactions. The transparency of the medium is determined by the opacity e, which describes the degree of absorption of electromagnetic radiation. The opacity is wavelength dependent. In an ionic crystal the absorption of infrared radiation is large. It alters the vibrational frequency, and thereby influences the ability of the crystal lattice to transport heat by conduction. Thus the effect can be taken into account by increasing the conductivity by an extra radiative amount, k_r, given by

$$k_r = \frac{16}{3}\frac{n^2\sigma}{e}T^3 \tag{4.44}$$

The T^3-dependence in this expression suggests that radiation might be more important than lattice conductivity in the hotter regions of the Earth. In fact other arguments lead to the conclusion that this is probably not the case in the upper mantle, because the effect of increasing temperature is partly offset by an increase in the opacity, e. The lower mantle is believed to have a high density of free electrons, which efficiently absorb radiation and raise the opacity. This may greatly reduce the efficacy of heat transfer by radiation in the mantle.

4.2.5 Sources of heat in the Earth

The interior of the Earth is losing heat via geothermal flux at a rate of about 4.4×10^{13} W, which amounts to 1.4×10^{21} J yr^{-1} (see Table 4.3). The heat is brought to the surface in different ways. The creation of new lithosphere at oceanic ridges releases the largest fraction of the thermal energy. A similar mechanism, the spreading of the sea-floor, releases heat in the marginal basins behind island arcs. Rising plumes of magma originating deep in the mantle bring heat to the surface where they break through the oceanic or continental lithosphere at

Table 4.4 *Estimates of radioactive heat production in selected rock types, based on heat production rates (from Rybach, 1976, 1988) and isotopic concentrations*

	Concentration [p.p.m. by weight]			Heat production [10^{-11} W kg^{-1}]			
Rock type	U	Th	K	U	Th	K	Total
Granite	4.6	18	33,000	43.8	46.1	11.5	101
Alkali basalt	0.75	2.5	12,000	7.1	6.4	4.2	18
Tholeiitic basalt	0.11	0.4	1,500	1.05	1.02	0.52	2.6
Peridotite, dunite	0.006	0.02	100	0.057	0.051	0.035	0.14
Chondrites	0.015	0.045	900	0.143	0.115	0.313	0.57
Continental crust	1.2	4.5	15,500	11.4	11.5	5.4	28
Mantle	0.025	0.087	70	0.238	0.223	0.024	0.49

"hotspots," characterized by intense localized volcanic activity. These important thermal fluxes are superposed on a background consisting of heat flowing into and through the lithosphere from deeper parts of the earth. There are two main sources of the internal heat. Part of it is probably due to the slow cooling of the Earth from an earlier hotter state; part is generated by the decay of long-lived radioactive isotopes.

The early thermal history of the Earth is obscure and a matter of some speculation. According to the cold accretion model of the formation of the planets (see Section 1.1.4), colliding bodies in a primordial cloud of dust and gas coalesced by self-gravitation. The gravitational collapse released energy that heated up the Earth. When the temperature reached the melting point of iron, a liquid core formed, incorporating also nickel and possibly sulfur or another light element associated with iron. The differentiation of a denser core and lighter mantle from an initially homogeneous fluid must have released further gravitational energy in the form of heat. The dissipation of Earth's initial heat still has an important effect on internal temperatures.

Energy released by short-lived radioactive isotopes may have contributed to the initial heating, but the short-lived isotopes would be consumed quite quickly. The heat generated by long-lived radioactive isotopes has been an important heat source during most of Earth's history. These isotopes separated into two fractions: some, associated with heavy elements, sank into the core; some, associated with lighter elements, accumulated in the crust. The present distribution of radiogenic sources within the differentiated Earth is uneven. The highest concentrations are in the rocks and minerals of the Earth's crust, while the concentrations in mantle and core materials are low, However, continuing generation of heat by radioactivity in the deep interior, though small, may influence internal temperatures.

4.2.5.1 Radioactive heat production

When a radioactive isotope decays, it emits energetic particles and γ-rays. The two particles that are important in radioactive heat production are α-particles and β-particles. The α-particles are equivalent to helium nuclei and are positively charged, while β-particles are electrons. In order to be a significant source of heat a radioactive isotope must have a half-life comparable to the age of the Earth, the energy of its decay must be fully converted to heat, and the isotope must be sufficiently abundant. The main isotopes that fulfil these conditions are ^{238}U, ^{235}U, ^{232}Th and ^{40}K. The isotope ^{235}U has a shorter half-life than ^{238}U (see Table 4.1) and releases more energy in its decay. In natural uranium the proportion of ^{238}U is 99.28%, that of ^{235}U is about 0.71%, and the rest is ^{234}U. The abundance of the radioactive isotope ^{40}K in natural potassium is only 0.01167%, but potassium is a very common element and its heat production is not negligible. The amounts of heat generated per second by these elements (in μW kg^{-1}) are: natural uranium, 95.2; thorium, 25.6; and natural potassium, 0.00348 (Rybach, 1976, 1988). The heat Q_r produced by radioactivity in a rock that has concentrations C_U, C_{Th} and C_K, respectively, of these elements is

$$Q_r = 95.2C_U + 25.6C_{Th} + 0.00348C_K \quad (4.45)$$

Rates of radioactive heat production computed with this equation are shown for some important rock types in Table 4.4. Chondritic meteorites, made up of silicate minerals like olivine and pyroxene, are often taken as a proxy for the initial composition of the mantle; likewise, the olivine-dominated rock dunite represents the ultramafic rocks of the upper mantle. It is apparent that very little heat is produced by radioactivity in the mantle or in the basaltic rocks that dominate the oceanic crust and lower continental crust. The greatest concentration of radiogenic heat sources is in the granitic rocks in the upper continental crust. Multiplying the radioactive heat production values in the last column of Table 4.4 by the rock density gives the radiogenic heat generated in a cubic meter of the rock, A. If we assume that all the heat generated in a rock layer of thickness D meters escapes vertically, the amount crossing a square meter at the surface per second (i.e., the radioactive component of the heat flow) is DA. For example, a layer of granite 1 km thick contributes about 3 mW m^{-2} to

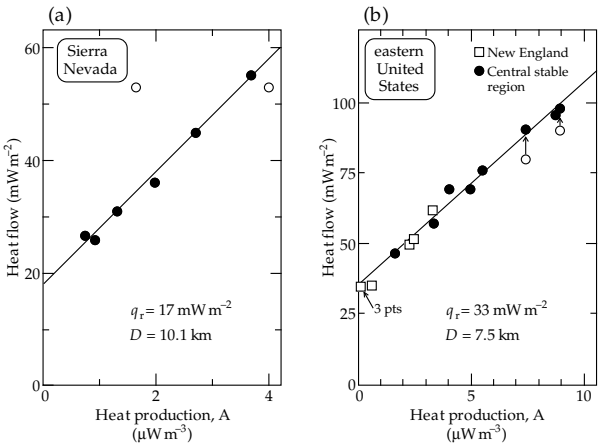

Fig. 4.16 Dependence of surface heat flux on radioactive heat generation in two heat-flow provinces: (a) Sierra Nevada, (b) eastern United States (data source: Roy *et al.*, 1968).

the continental heat flow. The figures suggest that the 10–20 km thick upper crust produces about one-half of the mean continental heat flow, which is 65 mW m^{-2}.

In fact the relative importance of radiogenic heat in the crust is variable from one region to another. A region in which the heat flow is linearly related to the heat produced by radioactivity is called a *heat-flow province*. Some examples of heat-flow provinces are Western Australia, the Superior Province in the Canadian Shield, and the Basin-and-Range Province in the western United States. As shown in Fig. 4.16 each province is characterized by a different linear relation between q and A, such that

$$q = q_r + DA \tag{4.46}$$

The parameters q_r and D typify the heat-flow province. The intercept of the straight line with the heat-flow axis, q_r, is called the *reduced heat flow*. This is the heat flow that would be observed in the province if there were no radiogenic crustal heat sources. It is due in part to the heat flowing from deeper regions of the Earth into the base of the crustal layer, and partly to cooling of the originally hotter upper crustal layer. Investigations in different heat-flow provinces show that the reduced heat flow averages about 55% of the mean measured heat flow in a province (Fig. 4.17).

The simplest interpretation of D is to regard it as a characteristic thickness of crust involved in radioactive heat production. This assumes that the radiogenic heat sources are distributed uniformly in a crustal slab of constant thickness, which is an unlikely situation. A more likely model is that the radioactive heat generation decreases with depth. Assuming an exponential decrease, the heat production $A(z)$ at depth z is related to the surface heat generation A_0 as

$$A(z) = A_0 e^{-z/D} \tag{4.47}$$

where D is a characteristic depth (the depth at which $A(z)$ has decreased to e^{-1} of its surface value). Integrating from

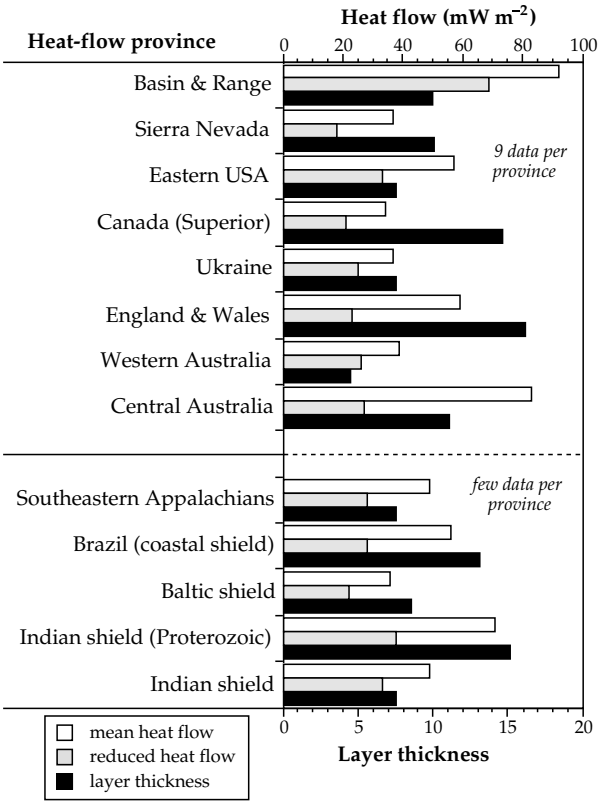

Fig. 4.17 Mean heat flow, reduced heat flow and characteristic thickness of the layer of radioactive heat production in several heat-flow provinces (data source: Vitorello and Pollack, 1980).

the surface to infinite depth gives the total radioactive heat production:

$$\int_0^\infty A(z)\,dz = A_0 \int_0^\infty e^{-z/D}\,dz = DA_0 \tag{4.48}$$

which is the same as for the uniform distribution. The infinite depth limit to the exponential distribution is obviously unrealistic. If the radiogenic sources are distributed in a layer of finite thickness s, the integration becomes

$$A_0 \int_0^s e^{-z/D}\,dz = DA_0(1 - e^{-s/D}) \tag{4.49}$$

If s is greater than three times D, this expression differs from DA_0 by less than 5%. The value of D estimated from studies of heat-flow and radioactive heat generation averages about 10 km, but varies from ~4 km to ~16 km (Fig. 4.17).

The three main sources of the Earth's surface heat flow are (i) heat flowing into the base of the lithosphere from the deeper mantle, (ii) heat lost by cooling of the lithosphere with time, and (iii) radiogenic heat production in the crust. The contributions are unequal and different in the oceans and continents (Table 4.5). The most obvious disparity is in the relative importance of lithospheric cooling and radioactivity. The lithosphere is hot when created at oceanic ridges and cools slowly as it ages. The

Table 4.5 *Approximate relative contributions (in %) of the main sources of heat flow in oceanic and continental lithosphere (from Bott, 1982)*

Heat source	Contribution to heat flow in: Continents [%]	Oceans [%]
Cooling of the lithosphere	20	85
Heat flow from below the lithosphere	25	10
Radiogenic heat:	55	5
upper crust	40	—
rest of lithosphere	15	—

loss of heat by lithospheric cooling is most pronounced in the oceanic crust, which moreover contains few radiogenic heat sources. In contrast, the older continental lithosphere has lost much of the early heat of formation, and the higher concentration of radioactive minerals increases the importance of radiogenic heat production. Regardless of its source, the passage of heat through the rigid outer layers takes place predominantly by conduction, although in special circumstances, such as the flow of magma in the crust or hydrothermal circulation near to oceanic ridges, convection also plays an important role.

4.2.6 The heat conduction equation

Jean Baptiste Joseph Fourier (1768–1830), a noted French mathematician and physicist, developed the theory of heat conduction in 1822. Here we consider the example of one-dimensional heat flow, which typifies many interesting problems involving the flow of heat in a single direction. The equation of heat conduction is most easily developed for this case, from which it can be readily extended to three-dimensional heat flow.

Consider the flow of heat in the negative direction of the z-axis through a small rectangular prism with sides $\mathrm{d}x$, $\mathrm{d}y$ and $\mathrm{d}z$ (see Fig. 4.15b). We will assume that there are no sources of heat inside the box. Let the amount of heat entering the prism at $(z+\mathrm{d}z)$ be Q_z. This is equal to the heat flow q_z multiplied by the area of the surface it flows across ($\mathrm{d}x\,\mathrm{d}y$) and by the duration of the flow ($\mathrm{d}t$). The heat leaving the box at z is $Q_z - \mathrm{d}Q_z$, which can be written $Q_z - (\mathrm{d}Q_z/\mathrm{d}z)\mathrm{d}z$. The increase in heat in the small box is the difference between these amounts:

$$\frac{\mathrm{d}Q_z}{\mathrm{d}z}\mathrm{d}z = \frac{\mathrm{d}q_z}{\mathrm{d}z}\mathrm{d}z\,(\mathrm{d}x\,\mathrm{d}y)\,\mathrm{d}t = k\frac{\mathrm{d}^2T}{\mathrm{d}z^2}\mathrm{d}V\,\mathrm{d}t \qquad (4.50)$$

where $\mathrm{d}V$ is the volume of the box ($\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$). Note that Q_z, q_z and T are all understood to decrease in the direction of flow, so we have substituted for q_z from Eq. (4.39) without using the negative sign. The heat increase in the box causes its temperature to rise by an amount ($\mathrm{d}T$), determined by the specific heat at constant pressure (c_p) and the mass of material (m) in the box. Using Eq. (4.25), we write

$$c_\mathrm{p} m\,\mathrm{d}T = c_\mathrm{p}\rho\,\mathrm{d}V\,\mathrm{d}T \qquad (4.51)$$

where ρ is the density of the material in the box. If we equate Eq. (4.50) and Eq. (4.51) for the amount of heat left in the box, we get the equation of heat conduction

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{k}{\rho c_\mathrm{p}}\frac{\mathrm{d}^2T}{\mathrm{d}z^2}$$

$$\frac{\partial T}{\partial t} = \kappa\frac{\partial^2 T}{\partial z^2} \qquad (4.52)$$

where κ ($= k/\rho c_\mathrm{p}$) is called the *thermal diffusivity*; it has the dimensions $\mathrm{m^2\,s^{-1}}$. The equation is written with partial differentials because the temperature T is a function of both time and position: $T = T(z,\ t)$. This just means that, on the one hand, the temperature at a certain position changes with time, and, on the other hand, the temperature at any given time varies with position in the body.

The same arguments can be applied to the components of heat flow through the box in the x- and y-directions. We obtain the three-dimensional heat flow equation (also called the *diffusion equation*):

$$\frac{\partial T}{\partial t} = \kappa\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right) \qquad (4.53)$$

The equation may be solved for any set of boundary conditions using the method of *separation of variables* (Box 4.1). Two important situations involving the flow of heat across the Earth's surface are the heating of Earth's surface by solar energy and the cooling of hot lithosphere. These can be handled to a first approximation as problems of one-dimensional heat conduction.

4.2.6.1 Penetration of external heat into the Earth

Solar energy is by far the greatest of the Earth's energy sources (Table 4.3). In order to determine the geothermal flux from the Earth's interior, it is important to understand the effects of solar energy that reaches the Earth's surface. Rocks at Earth's surface heat up during the day and cool down at night. The effect is not restricted to the immediate surface, but affects a volume of rock near the surface. Similarly, the mean surface temperature varies throughout the year with the changing seasons. The heat conduction equation allows us to estimate what depths are affected by these cyclic temperature variations.

Suppose that the surface temperature of the Earth varies cyclically with angular frequency ω, so that at time t it is equal to $T_0\cos\omega t$, where T_0 is the peak temperature during a cycle. The temperature at time t and depth z is obtained by solving the one-dimensional heat equation (Box 4.2) and is given by

Box 4.1: Method of separation of variables

The method of separation of variables is encountered in the solution of several geophysical problems. It may be illustrated by the case of one-dimensional heat flow, described by Eq. (4.52). The coordinates in this equation are independent, i.e., each can take any value regardless of the other. Assume that the temperature $T(z, t)$ can be written as the product of $Z(z)$, a function of position only, and $\theta(t)$, a function of time only:

$$T(z,t) = Z(z)\theta(t) \tag{1}$$

The actual solution for $T(z, t)$ usually does not have this final form, but once $Z(z)$ and $\theta(t)$ are known they can be combined to fit the boundary conditions. On substituting in Eq. (4.52) we get

$$Z\frac{\mathrm{d}\theta}{\mathrm{d}t} = \kappa\frac{\mathrm{d}^2 Z}{\mathrm{d}z^2}\theta \tag{2}$$

Note that we can use full differentials here because θ depends only on t and Z only on z. Dividing both sides by $Z\theta$ gives

$$\frac{1}{\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t} = \kappa\frac{1}{Z}\frac{\mathrm{d}^2 Z}{\mathrm{d}z^2} \tag{3}$$

Note that in Eq. (3) the left side depends only on t and the right side depends only on z, and that z and t are independent variables. On substituting a particular value for the time t the left side becomes a numerical constant. We have not restricted z, which varies independently. Equation (3) requires that the variable right side remains equal to the numerical constant for any value of z. Conversely, if we substitute a particular value for z, the right side becomes a (new) numerical constant and the variable left side must equal this constant for any t. The identity inherent in Eq. (3) implies that each side must be equal to the same *separation constant*. Let this constant be C. Then,

$$\begin{aligned}\frac{1}{\theta}\frac{\mathrm{d}\theta}{\mathrm{d}t} &= C\\ \kappa\frac{1}{Z}\frac{\mathrm{d}^2 Z}{\mathrm{d}z^2} &= C\end{aligned} \tag{4}$$

The value of the separation constant in a particular problem is determined by the boundary conditions of the problem.

$$T(z,t) = T_0\,\mathrm{e}^{-z/d}\cos\left(\omega t - \frac{z}{d}\right) \tag{4.54}$$

where $d = (2\kappa/\omega)^{1/2}$ is the *decay depth* of the temperature. At a depth of $5d$ the amplitude is less than 1% of the surface value and is effectively zero. Note that d depends inversely on the frequency, so long-period fluctuations penetrate more deeply than rapid fluctuations. This is illustrated by a quick comparison of the decay depths for daily and annual temperature variations in the same ground

$$\frac{d_{\text{annual}}}{d_{\text{daily}}} = \sqrt{\frac{(2\pi/1)}{(2\pi/365)}} = \sqrt{365} = 19.1 \tag{4.55}$$

i.e., the annual variation penetrates about 19 times the depth of the daily variation (Fig. 4.18). Moreover, the temperature at any depth varies with the same frequency ω as the surface variation, but it experiences a phase shift or delay, reaching its maximum value progressively later than the surface effect (Fig. 4.18).

As representative values for crustal rocks we take: density, $\rho = 2650\,\mathrm{kg\ m^{-3}}$; thermal conductivity, $k = 2.5\,\mathrm{W\,m^{-1}\,K^{-1}}$; and specific heat, $c_\mathrm{p} = 700\,\mathrm{J\,kg^{-1}\,K^{-1}}$. These give a thermal diffusivity, $\kappa = 1.25\times10^{-6}\,\mathrm{m^2\,s^{-1}}$. The daily temperature variation (period 86,400 s) has $\omega = 7.27\times10^{-5}\,\mathrm{s^{-1}}$, so its penetration depth is about 20 cm. The daily variation has negligible effect deeper than about a meter. Similarly, the annual temperature variation accompanying seasonal changes has a penetration depth of 3.8 m and is negligible deeper than about 19 m. Heat-flow measurements made within 20 m of the surface will be contaminated by the daily and annual variations of surface temperature. This is not a problem in the deep oceans, where the Sun's rays never reach the bottom, but it must be taken into account in continental heat-flow measurements. A serious effect is the role of the ice ages, which recur on a timescale of about 100,000 yr and have a penetration depth of several kilometers. The measured temperature gradient must be corrected appropriately.

4.2.6.2 Cooling of the oceanic lithosphere

A thermodynamic problem commonly encountered in geology is the sudden heating or cooling of a body. For example, molten lava intrudes cool host rocks as a dike or sill virtually instantaneously, but the heat is conducted to the adjacent rocks over a long period of time until it slowly dissipates. An important case in the context of plate tectonics is the cooling of the oceanic lithosphere, with its base at the temperature of the hot mantle, T_m, and its top in contact with cold ocean-bottom sea-water.

Assume that the fresh hot lithosphere is created at a ridge axis as a thin vertical prism with a uniform initial temperature equal to that of the hot mantle, T_m. Sea-floor spreading transports the lithosphere horizontally, so, for a constant spreading rate, distance from the ridge axis is proportional to the cooling time (t). Except in the immediate vicinity of a spreading ridge the heat loss may be assumed to be solely in the vertical ($-z$) direction through the surface at $z = 0$, which is in contact with cold ocean-bottom sea-water at a temperature of 0 °C. Although the lithosphere has finite vertical extent, the temperature in the cooling slab may be approximated as the one-dimensional cooling of a semi-infinite half-space extending to infinity in the z-direction. The error introduced by

Box 4.2: One-dimensional heat conduction

The surface-temperature variation $T_0 \cos \omega t$ can be expressed with the aid of complex numbers (Box 2.5) as the real part of $T_0\, e^{i\omega t}$. Let z represent the depth of a point below the surface. The heat conduction equation, Eq. (4.52), can be separated and written as two parts equal to the same constant. We want to match boundary conditions with the surface disturbance $T_0\, e^{i\omega t}$, so we write the separation constant as $i\omega$. Applying the method of separation of variables (Box 4.1) we get for the two parts of the solution

$$\frac{1}{\theta}\frac{d\theta}{dt} = i\omega \tag{1}$$

$$\kappa \frac{1}{Z}\frac{d^2Z}{dz^2} = i\omega \tag{2}$$

The solution of Eq. (1) gives the time dependence of the temperature as

$$\theta = \theta_0 e^{i\omega t} \tag{3}$$

In order to find the depth dependence, we can rewrite Eq. (2) as

$$\frac{d^2Z}{dz^2} - i\frac{\omega}{\kappa}Z = 0 \tag{4}$$

Writing $-n^2 = i\omega/\kappa$, this is equivalent to the harmonic equation

$$\frac{d^2Z}{dz^2} + n^2 Z = 0 \tag{5}$$

which has the solutions

$$Z = Z_0 e^{inz} \quad \text{and} \quad Z = Z_1 e^{-inz} \tag{6}$$

where

$$in = \sqrt{i\frac{\omega}{\kappa}} = \sqrt{\frac{\omega}{2\kappa}}(1+i) \tag{7}$$

We get two possible solutions for the depth variation:

$$Z = Z_1 e^{inz} = Z_1 e^{(\sqrt{\omega/2\kappa})(1+i)z} \quad \text{and}$$
$$Z = Z_0 e^{-inz} = Z_0 e^{-(\sqrt{\omega/2\kappa})(1+i)z} \tag{8}$$

The temperature must decrease with increasing depth z below the surface, so only the second solution is acceptable. Combining the solutions for θ and Z we get

$$T(z,t) = Z_0 e^{-(\sqrt{\omega/2\kappa})(1+i)z}\theta_0 e^{i\omega t} \tag{9}$$

$$T(z,t) = Z_0\theta_0 e^{-(\sqrt{\omega/2\kappa})z} e^{i(\omega t - (\sqrt{\omega/2\kappa})z)} \tag{10}$$

The surface-temperature variation $T_0 \cos \omega t$ was expressed as the real part of $T_0\, e^{i\omega t}$. Taking the real part of Eq. (10) we get for the temperature $T(z,t)$ at time t and depth z

$$T(z,t) = Z_0\theta_0 e^{-\sqrt{\omega/2\kappa}z}\cos\left(\omega t - \sqrt{\frac{\omega}{2\kappa}}z\right) \tag{11}$$

Writing $T_0 = Z_0\theta_0$ and $d = \sqrt{(2\kappa/\omega)}$ the equation reduces to

$$T(z,t) = T_0\, e^{-z/d}\cos\left(\omega t - \frac{z}{d}\right) \tag{12}$$

The parameter d is called the *decay depth* of the temperature. At this depth the amplitude of the temperature fluctuation is attenuated to 1/e of its value on the surface. Eq. (12) can also be written in the form

$$T(z,t) = T_0 e^{-z/d}\cos\omega(t - t_d) \tag{13}$$

where the phase difference, or *delay time*,

$$t_d = \frac{z}{\omega d} \tag{14}$$

represents the length of time by which the temperature at depth z lags behind the surface temperature.

this simplification is small, and the simple model can be used to estimate the heat flow from the cooling lithosphere with acceptable accuracy.

The one-dimensional cooling of a semi-infinite half-space is described in Appendix B. The solution with the stated boundary conditions involves the error function, which is described in Box 4.3. The error function (erf) and complementary error function (erfc) of the parameter x are defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}du$$
$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) \tag{4.56}$$

The shapes of these functions are shown in Fig. 4.19; their values for any particular value of x are obtained from tables, just like other statistical or trigonometric functions. The complementary error function sinks asymptotically to zero, and is effectively zero for $x \geq 2$. The temperature T at depth z and time t after the half-space starts to cool is given by

$$T = T_m \mathrm{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right)$$
$$= T_m \mathrm{erf}(\eta) \quad \text{with} \quad \eta = \frac{z}{2\sqrt{\kappa t}} \tag{4.57}$$

where T_m is the temperature of the hot mantle, κ is the thermal diffusivity of the half-space, and the surface ($z = 0$) is at the temperature of the ocean floor, which is taken to be 0 °C.

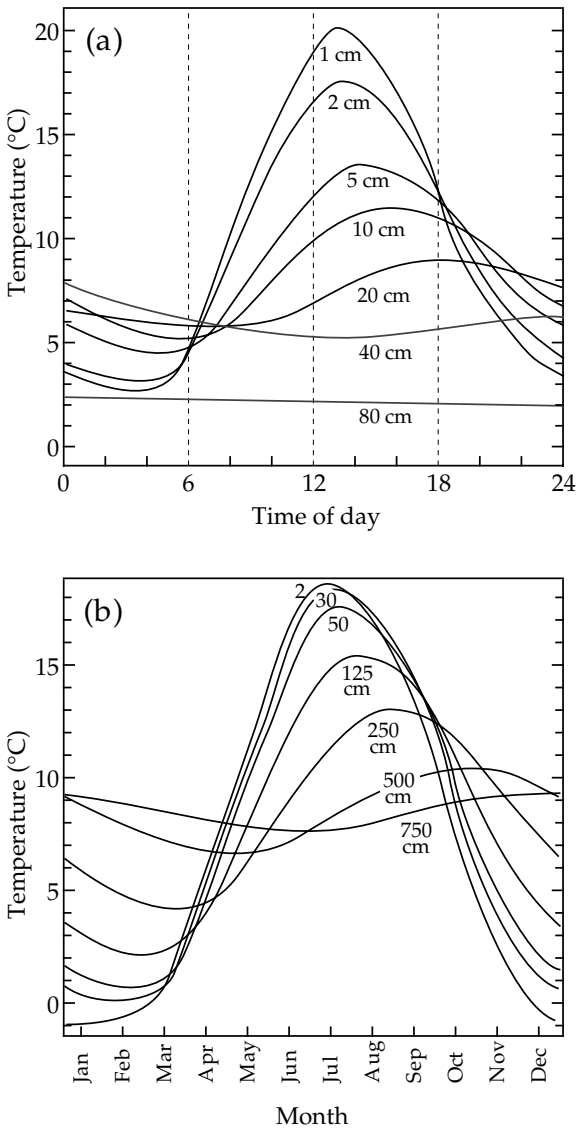

Fig. 4.18 Temperature variations at various depths in a sandy soil: (a) daily fluctuations, (b) annual (seasonal) variations.

The heat loss from the cooling lithosphere can be computed from the temperature distribution in Eq. (4.57). The heat flow is proportional to the temperature gradient (Eq. (4.39)). The heat flowing out of a semi-infinite half-space is thus obtained by differentiating Eq. (4.57) with respect to z:

$$\begin{aligned} q_z &= -k\frac{\partial T}{\partial z} = -k\frac{\partial \eta}{\partial z}\frac{\partial T}{\partial \eta} \\ &= -k\frac{1}{2\sqrt{\kappa t}}\frac{\partial}{\partial \eta}\{T_\mathrm{m}\,\mathrm{erf}(\eta)\} \\ &= -\frac{kT_\mathrm{m}}{2\sqrt{\kappa t}}\frac{\partial}{\partial \eta}\frac{2}{\sqrt{\pi}}\left\{\int_0^\eta \mathrm{e}^{-u^2}\mathrm{d}u\right\} \end{aligned} \tag{4.58}$$

which simplifies to

$$q_z = -\frac{kT_\mathrm{m}}{\sqrt{\pi\kappa t}}\mathrm{e}^{-\eta^2} \tag{4.59}$$

At the surface, $z = 0$, $\eta = 0$, and $\exp(-\eta^2) = 1$. The surface heat flow at time t is

$$q_z = -\frac{kT_\mathrm{m}}{\sqrt{\pi\kappa t}} \tag{4.60}$$

The negative sign here indicates that the heat flows upward, in the direction of decreasing z. The semi-infinite half-space is quite a good model for the cooling of oceanic lithosphere. The oceanic heat flow indeed varies with distance from an oceanic ridge as $1/\sqrt{t}$, where t is the corresponding age of the lithosphere. Models for the cooling of oceanic lithosphere are discussed further in Section 4.2.8.3.

4.2.7 Continental heat flow

The computation of heat flow at a locality requires two measurements. The thermal conductivities of a representative suite of samples of the local rocks are measured in the laboratory. The temperature gradient is measured in the field at the investigation site. At continental sites this is usually carried out in a borehole (Fig. 4.20). There are several ways of determining the temperature in the borehole. During commercial drilling the temperature of the drilling fluid can be measured as it returns to the surface. This gives a more or less continuous record, but is influenced strongly by the heat generated during the drilling. At times when drilling is interrupted, the bottom-hole temperature can be measured. Both of these methods give data of possibly commercial interest but they are too inaccurate for heat-flow determination.

In-hole measurements of temperature for heat-flow analyses are made by lowering a temperature-logging tool into the borehole and continuously logging the temperature during its descent. The circulation of drilling fluids redistributes heat in the hole, so it is necessary to allow some time after drilling has ceased for the hole to return to thermal equilibrium with the penetrated formations. The temperature of the water in the hole is taken to be the ambient temperature of the adjacent rocks, provided there are no convection currents.

The most common devices for measuring temperature are the platinum resistance thermometer and the thermistor. A thermistor is a ceramic solid-state device with an electrical resistance that is strongly dependent on temperature. Its resistance depends non-linearly on temperature, requiring accurate calibration, but the sensitivity of the device makes feasible the measurement of temperature differences of 0.001–0.01 K. The platinum resistance thermometer and the thermistor are used in two basic ways. In one method the sensor element constitutes an arm of a sensitive Wheatstone bridge, with which its resistance is measured directly. The other common method uses the thermal sensor as the resistive element in a tuned electrical circuit. The tuned frequency depends on the resistance of the sensor element, which is related in a known way to temperature.

Box 4.3: The error function

Errors may be of two types: systematic and random. A systematic error results, for example, when the measuring device is wrongly calibrated (e.g., if times are measured with a clock that runs "fast" or "slow"). Random errors occur naturally when a value is measured a large number of times. The observations will be distributed randomly about their mean value. Usually there will be a few large deviations from the mean and a lot of small deviations, and there will be as many negative as positive deviations. The scatter of the results can be described by the standard deviation σ of the measurements. Random errors are described by the *normal distribution*, which is often called a "bell curve" because of its shape (Fig. B4.3a). If the mean of the distribution of the paramer u is 0 and the standard deviation of the mean is σ, the normal distribution is described by the *probability density function*

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-(u/\sigma)^2/2} \tag{1}$$

The *standard normal distribution* is defined so that it has a mean of zero and standard deviation $\sigma = 1$. When integrated from $-\infty$ to $+\infty$, the area under this curve is

$$\int_{-\infty}^{\infty} f(u)\,du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\,du = 1 \tag{2}$$

The *error function* is closely related to the standard normal distribution. However, only positive values are considered, so the graph of the defining function is similar to the right half of the curve in Fig. B4.3a. It is given as

$$f(u) = \frac{2}{\sqrt{\pi}} e^{-u^2} \tag{3}$$

The area under this curve from the origin at $u = 0$ to the value $u = x$ (Fig. B4.3b) defines the *error function* erf(x):

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du \tag{4}$$

The *complementary error function*, erfc(x), is defined as

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-u^2} du \tag{5}$$

The value of erf(x) or erfc(x) for any particular value of x may be obtained from tables. Some useful properties of the error function and complementary error function are:

$$\mathrm{erf}(\infty) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-u^2} du = 1 \tag{6}$$

$$\frac{d}{dx}(\mathrm{erf}(x)) = \frac{2}{\sqrt{\pi}} e^{-x^2} \tag{7}$$

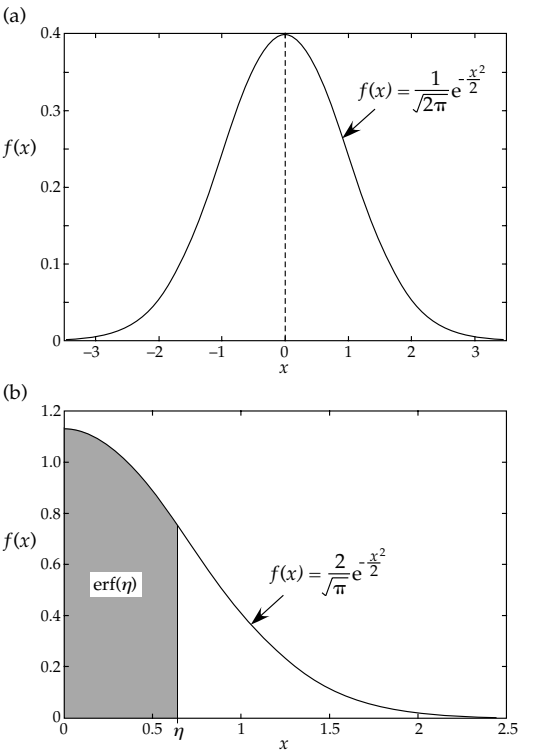

Fig. B4.3 (a) The normal distribution, and (b) the error function.

$$\begin{aligned}
\int_x^{\infty} \mathrm{erfc}(x)\,dx &= [x\cdot\mathrm{erfc}(x)]_x^{\infty} - \int_x^{\infty} x \frac{d}{dx}\Big(\mathrm{erfc}(x)\Big)\,dx \\
&= -x\,\mathrm{erfc}(x) - \int_x^{\infty} x\left(-\frac{2}{\sqrt{\pi}} e^{-x^2}\right)dx \\
&= \frac{2}{\sqrt{\pi}} \int_x^{\infty} x e^{-x^2} dx - x\,\mathrm{erfc}(x) \\
&= \frac{1}{\sqrt{\pi}}[-e^{-x^2}]_x^{\infty} - x\,\mathrm{erfc}(x) \\
&= \frac{e^{-x^2}}{\sqrt{\pi}} - x\,\mathrm{erfc}(x)
\end{aligned} \tag{8}$$

$$\int_0^{\infty} \mathrm{erfc}(x)\,dx = \frac{1}{\sqrt{\pi}} \tag{9}$$

Some values of erf(x) are listed in the following table.

x	erf(x)	x	erf(x)	x	erf(x)	x	erf(x)
0.05	0.05637	0.3	0.32863	0.6	0.60386	1.2	0.91031
0.1	0.11246	0.35	0.37938	0.7	0.67780	1.4	0.95229
0.15	0.16800	0.4	0.42839	0.8	0.74210	1.6	0.97635
0.2	0.22270	0.45	0.47548	0.9	0.79691	1.8	0.98909
0.25	0.27633	0.5	0.52050	1.0	0.84270	2.0	0.99532

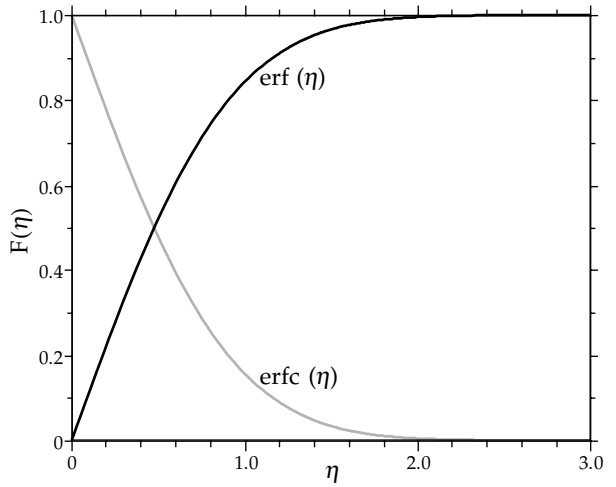

Fig. 4.19 The error function erf(η) and complementary error function erfc(η).

From the measured temperature distribution, the average temperature gradient is computed for a geological unit or a selected depth interval (Fig. 4.20). The gradient is then multiplied by the mean thermal conductivity of the rocks to obtain the interval (or formation) heat flow. Thermal conductivity can show large variations even between adjacent samples, so the *harmonic mean* of at least four samples is usually used. The interval heat-flow values may then be averaged to obtain the mean heat flow for the borehole (Table 4.6).

The computation of continental heat flow from borehole data requires the implementation of several corrections. An important assumption is that heat flow is only vertical. Well below the surface the isotherms (surfaces of constant temperature) are flat lying and the flow of heat (normal to the isotherms) is vertical. However, the surface of the Earth is also presumed to be locally isothermal (i.e., to have constant temperature) near to the borehole. The near-surface isotherms adapt to the topography (Fig. 4.21) so that the direction of heat flow is deflected and acquires a horizontal component, while the vertical temperature gradient is also modified. Consequently, heat flow measured in a borehole must be corrected for the effects of local topography.

The need for a topographic correction was recognized late in the nineteenth century. Further corrections must be applied for long-term effects such as the penetration of external heat related to cyclical climatic changes, for example the ice ages. Erosion, sedimentation and changes in the thermal conductivities of surface soils are other long-term effects that may require compensation.

4.2.7.1 Reconstruction of ground surface temperature changes from borehole temperature profiles

The main variation of temperature with depth in the Earth is the geothermal gradient related to the outflow of heat from the Earth's deep interior. Over limited depth intervals the temperature profile is linear, its slope varying by region with local conditions. However, changes in temperature at Earth's surface affect the subsurface temperature distribution close to the surface (Section 4.2.6.1). Rapid temperature changes have shallow penetration, but slow changes can modify subsurface temperatures well below the surface. For example, daily variations do not reach below about a meter, but temperature variations over a century may extend to about 200 m depth.

Table 4.6 *Computation of heat flow from temperature measurements in WSR.1, a 570 m deep borehole, and thermal conductivity measurements on cored samples (after Powell* et al.*, 1988)*

Depth interval [m]	Temperature gradient [mK m^{-1}]	Thermal conductivity [W m^{-1} K^{-1}]	Interval heat flow [mW m^{-2}]
45–105	15.0	3.96	60
105–245	18.0	3.43	62
245–320	24.8	2.75	68
320–455	16.0	4.18	67
455–515	17.2	4.20	72
515–575	16.5	3.86	64
Mean heat flow = 65 mW m^{-2}			

The temperature profiles in boreholes from eastern Canada (Fig. 4.22) show effects of surface temperature variations above depths of 180–250 m, below which the regular geothermal gradient is evident. The curved upper segment contains information about the history of temperature changes at the surface. This history can be retrieved by inversion of the borehole temperature data, which is a complex and sophisticated process. A large database of borehole temperature measurements, used for the determination of the global heat flow pattern (Section 4.2.8.2), has been analyzed to obtain the variation of mean surface temperature since 1500 (Fig. 4.23). Instrumental measurements of air temperature since 1860 agree well with the long-term surface temperatures estimated from the borehole data. The results show a consistent increase in surface temperature over the past five centuries.

4.2.7.2 Variation of continental heat flow with age

Many processes contribute to continental heat flow. Apart from the heat generated by radioactive decay, the most important sources are those related to tectonic events. During an orogenic episode various phenomena may introduce heat into the continental crust. Rocks may be deformed and metamorphosed in areas of continental collision. In extensional regions the crust may be thinned, with intrusion of magma. Uplift and erosion of elevated areas and deposition in sedimentary basins also affect the surface heat flow. After a tectonic event

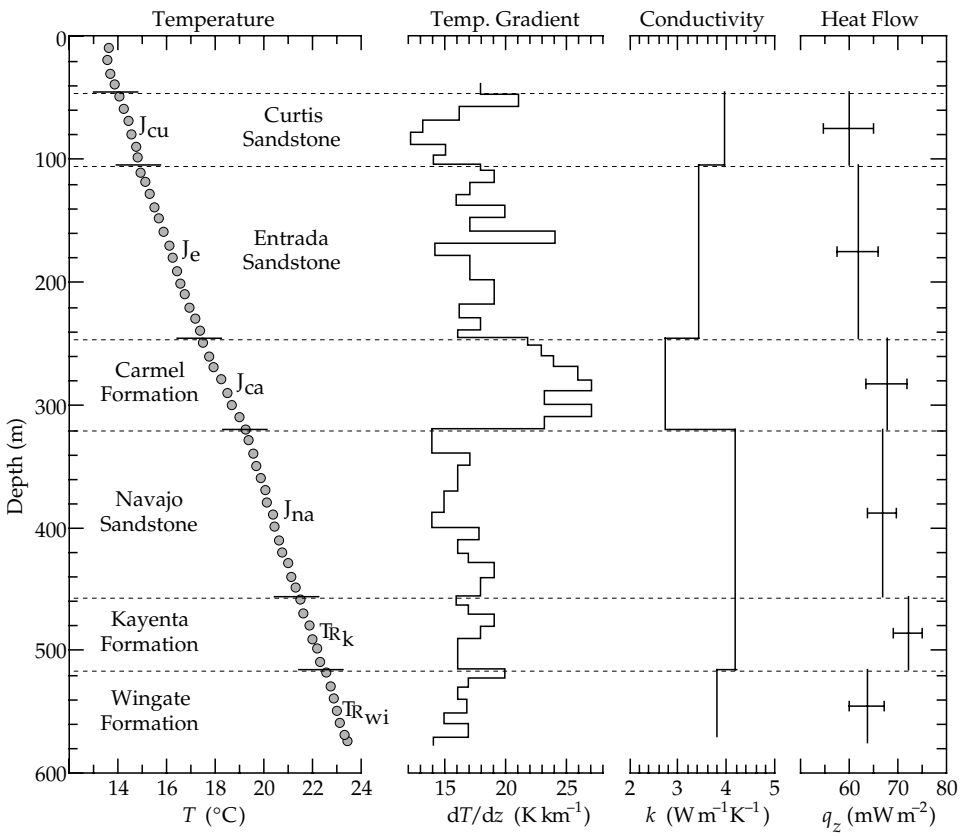

Fig. 4.20 Computation of heat flow by the interval method for geothermal data from drillhole WSR.1 on the Colorado Plateau of the western USA; horizontal bars show standard deviations of measurements in each depth interval (after Powell *et al.*, 1988; based on data from Bodell and Chapman, 1982).

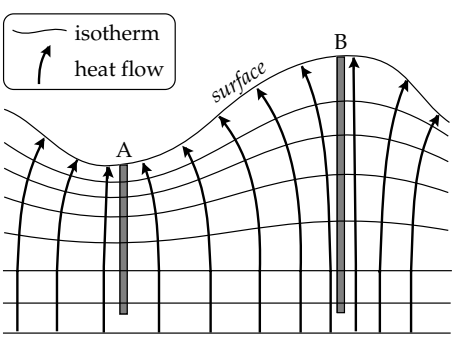

Fig. 4.21 Schematic effect of surface topography on isotherms (solid lines) and the direction of heat flow (arrows).

convective cooling takes place efficiently in circulating fluids, while some excess heat is lost by conductive cooling. Consequently, the variation of continental heat flow with time is best understood in terms of the tectonothermal age, which is the age of the last tectonic or magmatic event at a measurement site. The continental heat-flow values comprise a broad spectrum. Even when grouped in broad age categories there is a large degree of overlap (Fig. 4.24). The greatest scatter is seen in the youngest regions. The mean heat flow decreases with increasing crustal or tectonothermal age (Fig. 4.25), falling from 70–80 mW m^{-2} in young provinces to a steady-state value of 40–50 mW m^{-2} in Precambrian regions older than ~800 Ma .

4.2.7.3 Heat transfer through porous crustal rocks

The transfer of heat through the continental crust takes place not only by conduction but also by advection. A sediment or rock is composed of mineral grains closely packed in contact with each other. The pore spaces between the grains can represent an appreciable fraction of the total volume of a rock sample. This fraction, sometimes expressed as a percentage, is the *porosity* ϕ of the rock. A highly porous rock, for example, may have a porosity of 0.3, or 30%, implying that only 70% of the rock is solid mineral. The porosity depends on how the mineral grains are arranged, how well they are cemented, and on their degree of sorting. Well sorted sediments have fairly uniform grain and pore sizes; in poorly sorted sediments, there is a range of grain sizes, so the finer grains may "block" the voids between the larger grains, reducing the porosity. Igneous and metamorphic rocks often contain cracks and fissures, which, if sufficiently numerous, may give these rocks a low porosity. The degree to which the pore spaces are connected with each other determines the *permeability* of the rock, which is its ability to transmit fluids such as water and petroleum.

Permeability is defined by an empirical relationship observed in 1856 by Henry Darcy, a French hydraulic engineer. He observed that the volume of fluid per second crossing a surface was proportional to the area of the surface (A) and the gradient of the hydraulic pressure head (dp/dx) driving the flow, assumed here to be in the x-direction, and inversely to the viscosity (μ) of the liquid.

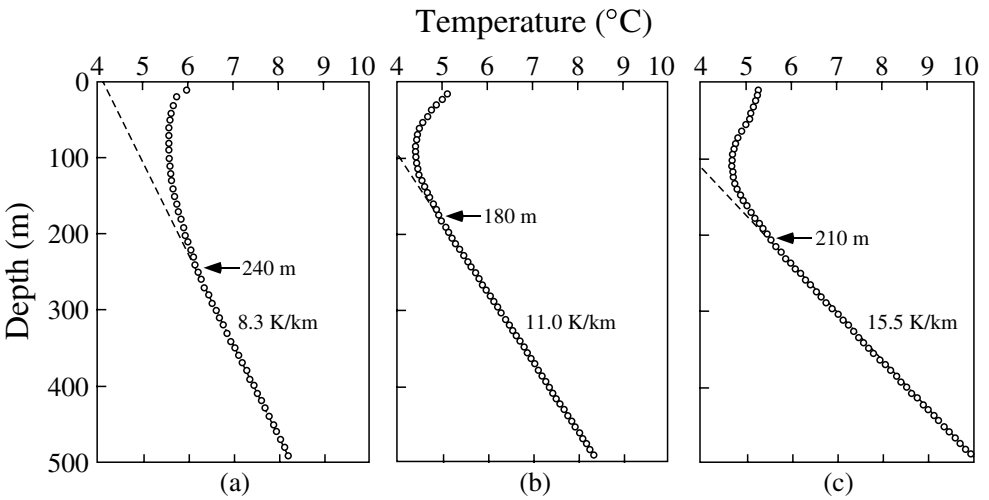

Fig. 4.22 Temperature–depth profiles in three boreholes in eastern Canada. The linear segment in the deeper part of each borehole is the local geothermal gradient. Climatically induced variations in the ground surface temperature result in the curved segment superposed on the linear record in approximately the upper 200 meters of each profile (after Pollack and Huang, 2000).

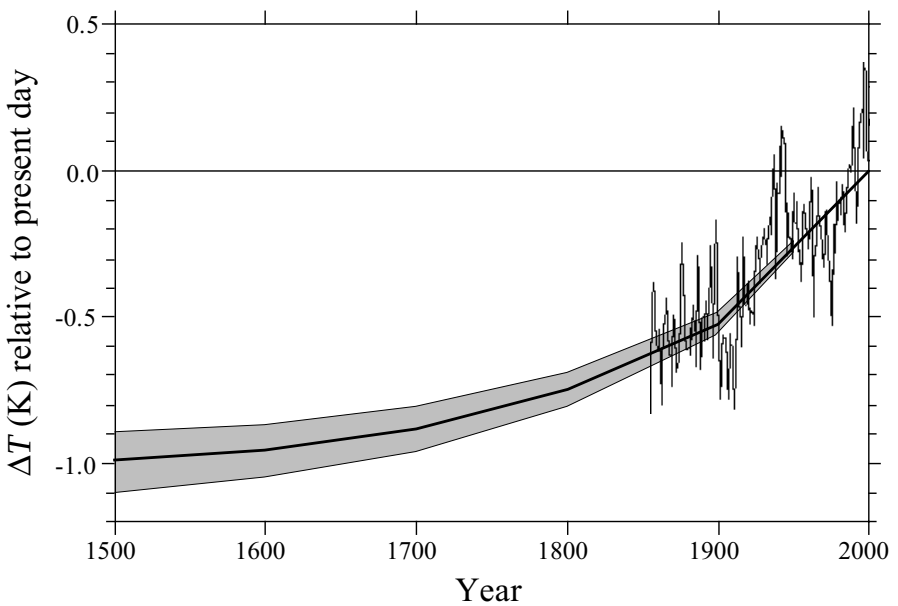

Fig. 4.23 History of surface temperature change (with ±1 standard error, shaded area) inferred for the past 500 years from a global database of borehole temperature measurements. The superposed signal since 1860 is a 5-year running mean of the globally averaged instrumental record of surface air temperature (after Pollack and Huang, 2000).

If the velocity of the flow is v, then in one second the volume crossing a surface A is (vA), so we have

$$(vA) \propto \frac{A}{\mu}\frac{\mathrm{d}p}{\mathrm{d}x} = -k\frac{A}{\mu}\frac{\mathrm{d}p}{\mathrm{d}x}$$

$$v = -\frac{k}{\mu}\frac{\mathrm{d}p}{\mathrm{d}x} \qquad (4.61)$$

$$v = -K\frac{\mathrm{d}p}{\mathrm{d}x}$$

This equation, in which $K = k/\mu$ defines the *hydraulic conductivity*, is known as Darcy's law. Its derivation and form are analogous to the law of heat conduction (Eq. (4.38), Fig. 4.15) and Ohm's law for electrical currents (Eq. (4.71), Fig. 4.40). The negative sign in the equation indicates that the flow is in the direction of decreasing pressure. The constant k is the permeability, which has dimensions m^2. However, the unit of permeability used in hydraulic engineering is called the *darcy* (equivalent to 9.87×10^{-13} m^2) or, more practically, the *millidarcy* (md). For example, an approximate range for the permeability of gravel is 10^5–10^8 md, that of sandstone is 1−100 md, and that of granite is 10^{-3}–10^{-5} md.

The ability of fluids to flow through crustal rocks enables them to transmit heat. In this case the process of heat transfer is not by convection, because the fluid motion is not driven by temperature differences but by the pressure gradient. The heat transfer "piggybacks" on the fluid motion, and the process is called *advection* (Section 4.2.4.2). The motion of water through the continental crust provides a source of geothermal energy, which can be tapped in several interesting ways for commercial purposes.

4.2.8 Oceanic heat flow

Whereas the mean altitude of the continents is only 840 m above sea-level, the mean depth of the oceans is 3880 m; the abyssal plains in large ocean basins are 5–6 km deep.

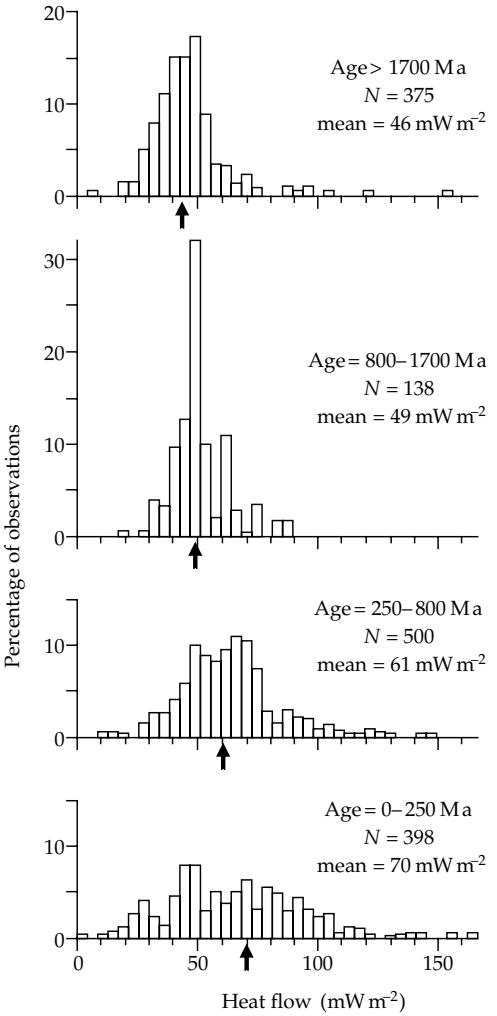

Fig. 4.24 Histograms of continental heat flow for four different age provinces (after Sclater *et al.*, 1981).

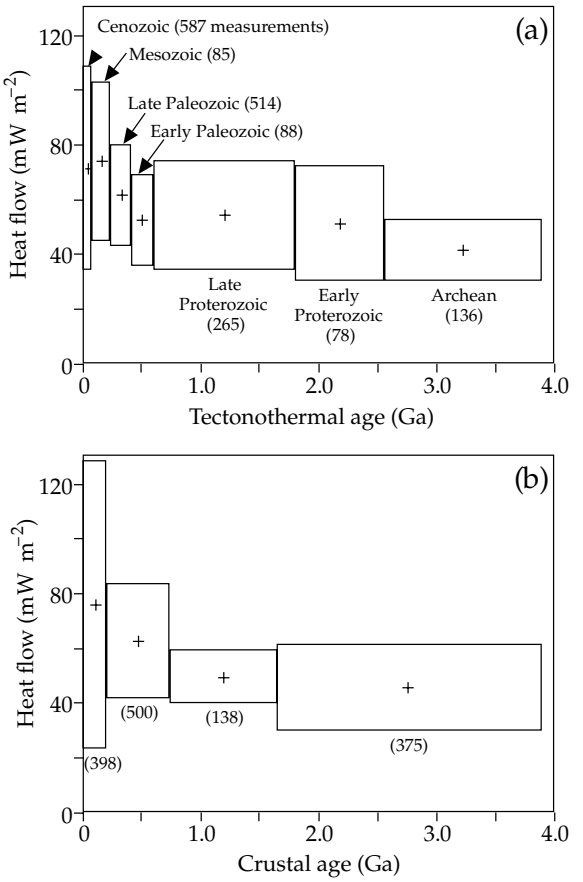

Fig. 4.25 Continental heat-flow data averaged (a) by tectonothermal age, defined as the age of the last major tectonic or magmatic event (based on data from Vitorello and Pollack, 1980), and (b) by radiometric crustal age (after Sclater *et al.*, 1980). The width of each box shows the age range of the data; the height represents one standard deviation on each side of the mean heat flow indicated by the cross at the center of each box. Numbers indicate the quantity of data for each box.

At these depths extra-terrestrial heat sources have no effect on heat-flow measurements and the flatness of the ocean bottom (except near ridge systems or seamounts) obviates the need for topographic corrections. Measuring the heat flow through the ocean bottom presents technical difficulties that were overcome with the development of the Ewing piston corer. This device, intended for taking long cores of marine sediment from the ocean floor, enables *in situ* measurement of the temperature gradient. It consists of a heavily weighted, hollow sampling pipe (Fig. 4.26a), commonly about 10 m long although in special cases cores over 20 m in length have been taken (very long coring pipes tend to bend before they reach maximum penetration). A plunger inside the pipe is displaced by sediment during coring and makes a seal with the sediment surface, so that sample loss and core deformation are minimized when the core is withdrawn from the ocean floor. Thermistors are mounted on short arms a few centimeters from the body of the pipe, and the temperatures are recorded in a water-tight casement. The instrument is lowered from a surface ship until a free-dangling trigger-weight makes contact with the bottom (Fig. 4.26b). This releases the corer, which falls freely and is driven into the sediment by the one-ton lead weight. The friction accompanying this process generates heat, but the ambient temperatures in the sediments can be measured and recorded before the heat reaches the offset sensors (Fig. 4.26c). The sediment-filled corer is hauled back on board the ship, where the thermal conductivity of the sediment can be determined. The recovered core is used for paleontological, sedimentological, geochemical, magnetostratigraphic and other scientific analyses.

Special probes have been devised explicitly for *in situ* measurement of heat flow. They consist of two parallel tubes about 3–10 m in length and 5 cm apart. One tube is about 5–10 cm in diameter and provides strength; the other, about 1 cm in diameter, is oil filled and contains arrays of thermistors. After penetration of the ocean-bottom sediments, as described for the Ewing corer, the equilibrium temperature gradient is measured. A known electrical current, either constant in value or pulsed, is then passed along a heating wire and the temperature response is recorded. The observations allow the thermal

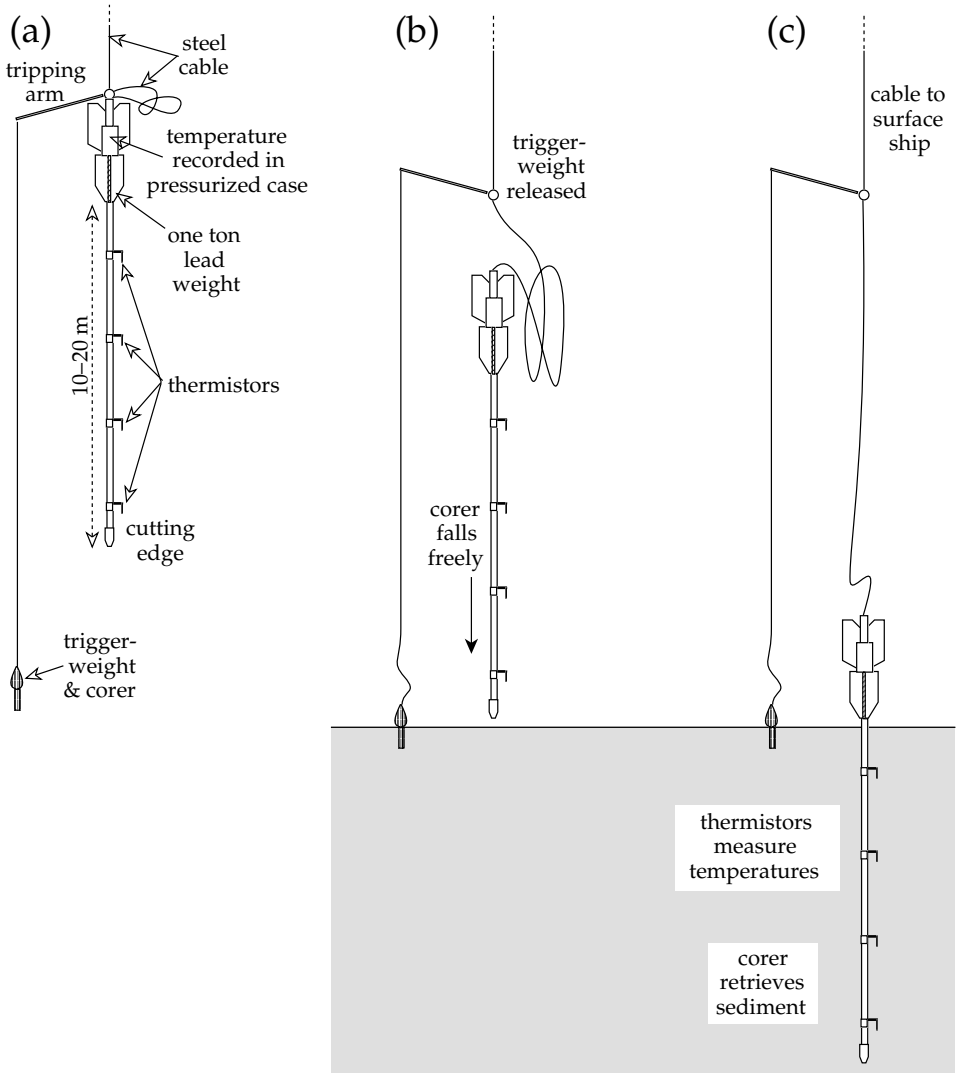

Fig. 4.26 Method of measuring oceanic heat flow and recovering samples of marine sediments: (a) a coring device is lowered by cable to the sea-floor, (b) when a trigger-weight contacts the bottom, the corer falls freely, and (c) temperature measurements are made in the ocean floor and the sediment-filled corer is recovered to the surface ship.

conductivity of the sediment to be found. In this way a complete determination of heat flow is obtained without having to recover the contents of the corer.

4.2.8.1 Variation of oceanic heat flow and depth with lithospheric age

The most striking feature of oceanic heat flow is the strong relationship between the heat flow and distance from the axis of an oceanic ridge. The heat flow is highest near to the ridge axis and decreases with increasing distance from it. For a uniform sea-floor spreading rate the age of the oceanic crust (and lithosphere) is proportional to the distance from the ridge axis, and so the heat flow decreases with increasing age (Fig. 4.27). The lithospheric plate accretes at the spreading center, and as the hot material is transported away from the ridge crest it gradually cools. Model calculations for the temperature in the cooling plate are discussed in the next section: they all predict that the heat flow q caused by cooling of the plate decreases with age t as $1/\sqrt{t}$, when the age of the plate is less than about 55–70 Ma. Older lithosphere cools slightly less rapidly. Currently the decrease in heat flow with age is best explained by a global model called the Global Depth and Heat Flow model (GDH1). The model predicts the following relationships between heat flow (q, mW m^{-2}) and age (t, Ma):

$$q = \frac{510}{\sqrt{t}} \qquad (t \leq 55 \text{ Ma})$$

$$q = q_s[1 + 2\exp(-\kappa\pi^2 t/a^2)] \quad (t > 55 \text{ Ma}) \qquad (4.62)$$

$$= 48 + 96\exp(-0.0278t)$$

Here q_s is the asymptotic heat flow, to which the heat flow decreases over very old oceanic crust (≈ 48 mW m^{-2}), a is the asymptotic thickness of old oceanic lithosphere ($\approx$95 km), and κ is its thermal diffusivity ($\approx 0.8 \times 10^{-6}$ m^2 s^{-1}).

Close to a ridge axis the measured heat flow is unpredictable: extremely high values and very low values have been recorded. Over young lithosphere the observed heat flow is systematically less than the values predicted by

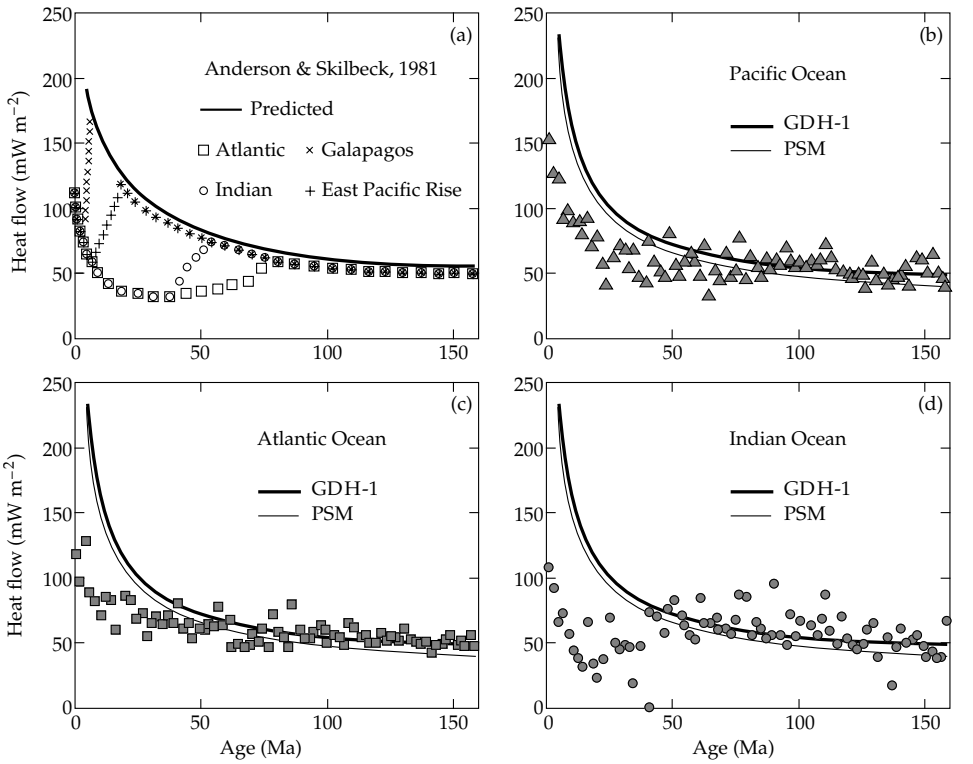

Fig. 4.27 Comparison of observed and predicted heat flow as a function of age of oceanic lithosphere. (a) Schematic summary for all oceans, showing the influence of hydrothermal heat flow at the ocean ridges (after Anderson and Skilbeck, 1981). Comparisons with the reference cooling models PSM (Parsons and Sclater, 1977) and GDH1 (Stein and Stein, 1992) for (b) the Pacific, (c) Atlantic and (d) Indian oceans.

cooling models (Fig. 4.27). The divergence is related to the process of accretion of the new lithosphere. At a ridge crest magma erupts in a narrow zone through feeder dikes and/or supplies horizontal lava flows. Very hot material is brought in contact with sea water, which cools and fractures the fresh rock. The water is in turn heated rapidly and a hydrothermal circulation is set up, which transports heat out of the lithosphere by convection. The eruption of hot hydrothermal currents has been observed directly from manned submersibles in the axial zones of oceanic ridges. The expeditions witnessed strong outpourings of mineral-rich hot water (called "black smokers" and "white smokers") in the narrow axial rift valley. The heat output of these vents is high: the power associated with a single vent has been estimated to be about 200 MW.

About 30% of the hydrothermal circulation takes place very near to the ridge axis through crust younger than 1 Ma. The rest is due to off-ridge circulation, which is possible because the fractured crust is still permeable to sea-water at large distances from the ridge axis. As it moves away from the ridge, sedimentation covers the basement with a progressively thicker layer of low permeability sediments, inhibiting the convective heat loss. The hydrothermal circulation eventually ceases, perhaps because it is sealed by the thick sediment cover, but probably also because the cracks and pore spaces in the crust become closed with increasing age. This is estimated to take place by about 55–70 Ma, because for greater ages the observed decrease in heat flow is close to that predicted by plate cooling models. The hydrothermal circulation in oceanic crust is an important part of the Earth's heat loss. It accounts for about a third of the total oceanic heat flow, and a quarter of the global heat flow.

The free-air gravity anomaly over an oceanic ridge system is generally small and related to ocean-bottom topography (Section 2.6.4.3), which suggests that the ridge system is isostatically compensated. As hot material injected at the ridge crest cools, its volume contracts and its density increases. To maintain isostatic equilibrium a vertical column sinks into the supporting substratum as it cools. Consequently, the depth of the ocean floor (the top surface of the column) is expected to increase with age of the lithosphere. The cooling half-space model predicts an increase in depth proportional to $\sqrt{t}$, where t is the age of the lithosphere, and this is observed up to an age of about 80 Ma (Fig. 4.28). However, the square-root relationship is not the best fit to the observations. Other cooling models fit the observations more satisfactorily, although the differences from one model to another are small. Beyond 20 Ma the data are better fitted by an exponential decay. The optimum relationships between depth (d, m) and age (t, Ma) can be written

$$\begin{aligned} d &= 2600 + 365\sqrt{t} && (t < 20 \text{ Ma}) \\ d &= d_r + [1 - (8/\pi^2)\exp(-k\pi^2 t/a^2)] && (t \geq 20 \text{ Ma}) \\ &= 5651 - 2473\exp(-0.0278t) \end{aligned} \quad (4.63)$$

where d_r is the mean depth of the ocean floor at ridge crests, d_s is the asymptotic subsidence of old lithosphere and the other parameters are as before.

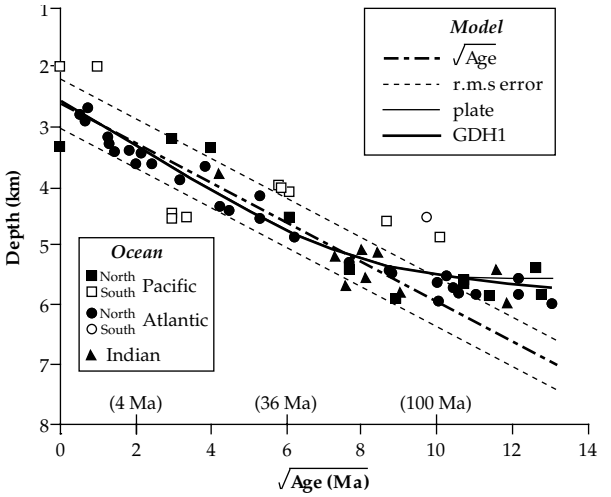

Fig. 4.28 Relationship between mean ocean depth and the square root of age for the Atlantic, Pacific and Indian oceans, compared with theoretical curves for different models of plate structure (after Johnson and Carlson, 1992).

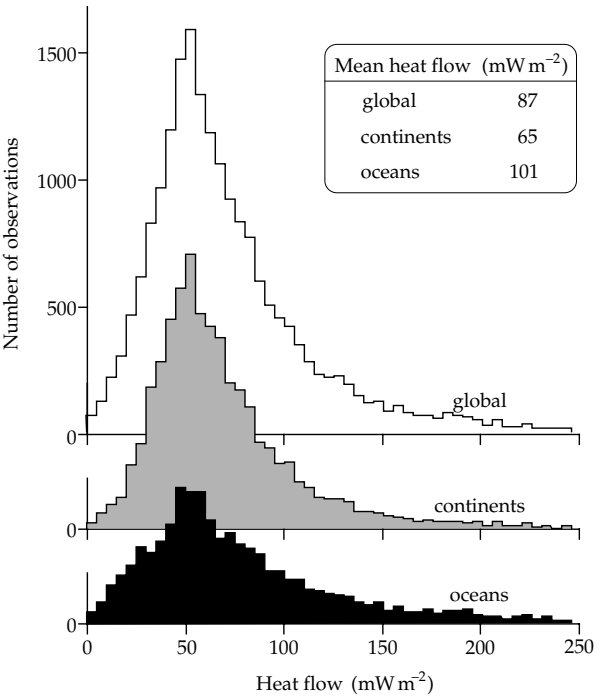

Fig. 4.29 Histograms of continental, oceanic and global heat-flow values (after Pollack *et al.*, 1993).

4.2.8.2 Global heat flow

Oceanic heat flow has been measured routinely in oceanographic surveys since the 1950s and *in situ* profiles have been made since the 1970s. In contrast to the measurement of continental heat flow it is not necessary to have an available (and usually expensive) drillhole. However, the areal coverage of the oceans by heat-flow measurements is uneven. A large area in the North Pacific Ocean is still unsurveyed, and most of the oceanic areas south of about latitude 35°S (the approximate latitude of Cape Town or Buenos Aires) are unsurveyed or only sparsely covered. The uneven data distribution is dense along the tracks of research vessels and absent or meager between them. The sites of measured continental heat flow are even more irregularly distributed. Antarctica, most of the interiors of Africa and South America, and large expanses of Asia are either devoid of heat-flow data or are represented by only a few sites.

In recent years a global data set of heat-flow values has been assembled, representing 20,201 heat-flow sites. The data set is almost equally divided between observations on land (10,337 sites) and in the oceans (9,864 sites). Histograms of the heat-flow values are spread over a wide range for each domain (Fig. 4.29). The distributions have similar characteristics, extending from very low, almost zero values to more than 200 mW m^{-2}. The high values on the continents are from volcanic and tectonically active regions, while the highest values in the oceans are found near to the axes of oceanic ridges. Both on the continents (Fig. 4.25) and in the oceans (Fig. 4.27), heat flow varies with crustal age. To determine global heat-flow statistics, the fraction of the Earth's surface area having a given age is multiplied by the mean heat flow measured for that age domain. The weighted sum gives a mean heat flow of 65 mW m^{-2} for the continental data set. The oceanic data must be corrected for hydrothermal circulation in young crust; the areally weighted mean heat flow is then 101 mW m^{-2} for the oceanic data set. The oceans cover 60.6% and the continents 39.4% of the Earth's surface, the latter figure including 9.1% for the continental shelves and other submerged continental crust. The weighted global mean heat flow is 87 mW m^{-2}. Multiplying by the Earth's surface area, the estimated global heat loss is found to be 4.42×10^{13} W (equivalent to an annual heat loss of 1.4×10^{21} J). About 70% of the heat is lost through the oceans and 30% through the continents.

The heat-flow values in both continental and oceanic domains are found to depend on crustal age and geological characteristics. These relationships make it possible to create a map of global heat flow that allows for the uneven distribution of actual measurements (Fig. 4.30). The procedure in creating this map was as follows. First, the Earth's surface was divided into 21 geological domains, of which 12 are in the oceans and 9 on the continents. Next, relationships between heat flow and age were used to associate a representative heat flow with each domain (Table 4.7). This made it possible to estimate heat flow for regions that have no measurement sites. Allowance was also made for the loss of heat by hydrothermal circulation near to ridge systems. The surface of the globe was next divided into a grid of 1° elements (i.e., each element measures 1° × 1°), and the mean heat flow through each element was estimated. This gave a complete data set (partly observed and partly synthesized) covering the entire globe. The gridded data were fitted by spherical harmonic functions (as in the representation of the geoid, Section 2.4.5) up to degree and

Table 4.7 *Mean heat-flow values for the oceans and continents, based on measurements at 20,201 sites (after Pollack* et al.*, 1993)*

The oceanic heat-flow values in italics are corrected for hydrothermal circulation according to the model of Stein and Stein (1992).

Description	Number of sites	Area of Earth [%]	Heat flow [mW m^{-2}]
Oceans			
Quaternary	415	1.2	*806*
Pliocene	712	2.4	*286*
Miocene	1,211	9.2	*142*
Oligocene	593	7.7	*93*
Eocene	691	7.8	*75*
Paleocene	205	3.9	*65*
Late Cretaceous	359	6.9	60
Middle Cretaceous	695	11.2	54
Early Cretaceous	331	4.3	51
Late Jurassic	295	3.8	49
Cenozoic undifferentiated	846	2.2	89
Mesozoic undifferentiated	599	0.2	45
All oceanic data	6,952	60.6	101
Continents			
Continental shelf regions	295	9.1	78
Cenozoic: igneous	3,705	1.1	97
sedimentary and metamorphic	2,912	8.1	64
Mesozoic: igneous	1,591	1.6	64
sedimentary and metamorphic	1,310	4.5	64
Paleozoic: igneous	1,810	0.4	61
sedimentary and metamorphic	403	5.9	58
Proterozoic	260	6.2	58
Archean	963	2.5	52
All continental data	13,249	39.4	65

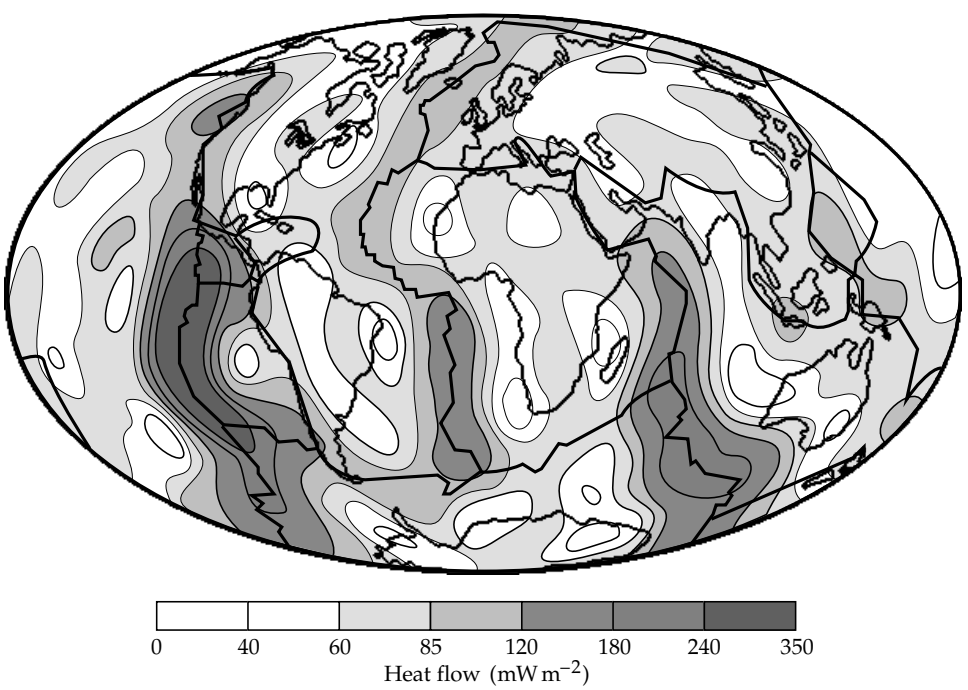

Fig. 4.30 Global distribution of heat flow (mW m^{-2}). The contours show a degree and order 12 spherical harmonic representation of the global heat flow based on direct measurements and empirical estimators for regions without data (after Pollack *et al.*, 1993).

order 12. The results of the analysis were used to compute smooth contours of equal heat flow, which were then plotted as the global heat-flow map (Fig. 4.30). If the global mean value is subtracted, the Earth's surface can be divided into regions with above-average and below-average heat flow, respectively (Fig. 4.31). The regions with above-average heat flow are notably associated with the oceanic ridge systems. About half of the Earth's heat is lost by the

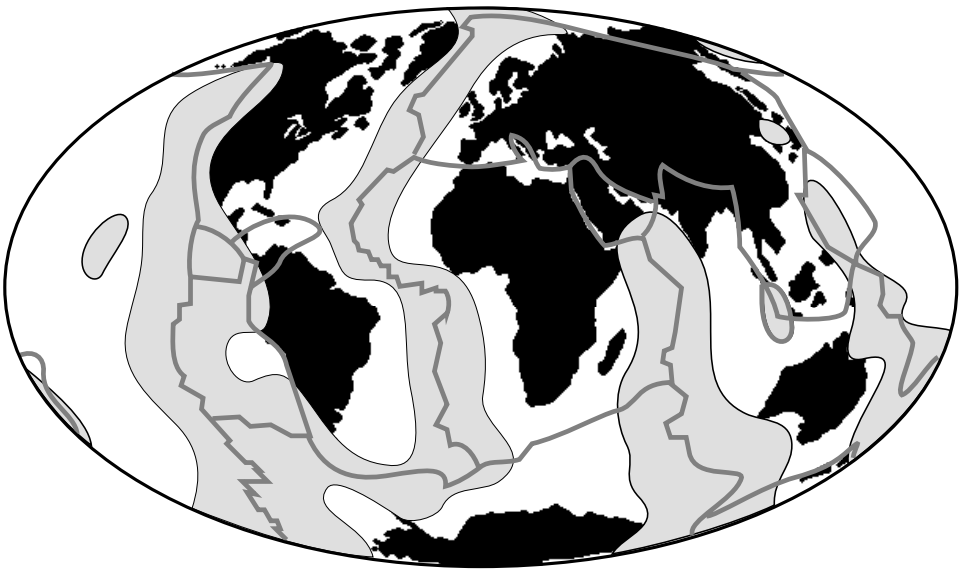

Fig. 4.31 Geographic regions where the heat flow is higher (lighter shaded) and lower (unshaded) than the global mean heat flow; lines (darker shaded) mark positions of plate boundaries (after Pollack *et al.*, 1993).

cooling of oceanic lithosphere of Cenozoic age (younger than 65 Ma).

One must keep in mind that this global model is based on a mixture of actual heat-flow measurements in regions where they are available, and estimated values in inaccessible regions. Moreover, the measured data near ocean ridges are replaced with values predicted by cooling models to compensate for the known loss of heat by hydrothermal circulation. Nevertheless, these global heat-flow maps are the best available representations of the geographical pattern and flux of the heat flowing out of the Earth's interior. Although details may eventually need modification, the main features are not in doubt.

4.2.8.3 Models for the cooling of oceanic lithosphere

The variations of heat flow and ocean depth with time constrain the possible thermal models for cooling of the lithosphere in different ways. The predicted heat flow is contingent on the temperature gradient in a model, but the oceanic depth is defined by the vertical distribution of density, which, in turn, depends on the volume coefficient of expansion and the temperature profile in the plate. Thus, oceanic bathymetry depends on the temperature integrated over depth.

Several cooling models have been proposed, all of which satisfy the decrease in heat flow and increase in ocean depth with age. The simplest model represents the cooling lithosphere as a semi-infinite half-space (Fig. 4.32a). Initially, the temperature inside the half-space is uniform and higher than on its upper surface, which is maintained at the temperature of cold ocean-bottom seawater. As long as the lithosphere is thin – which it is near the ridge – horizontal heat conduction can be neglected. The heat flow in the uniform half-space is vertical, along the z-axis, and is equivalent to the one-dimensional flow in a thin vertical column (Fig. 4.33a). The spreading

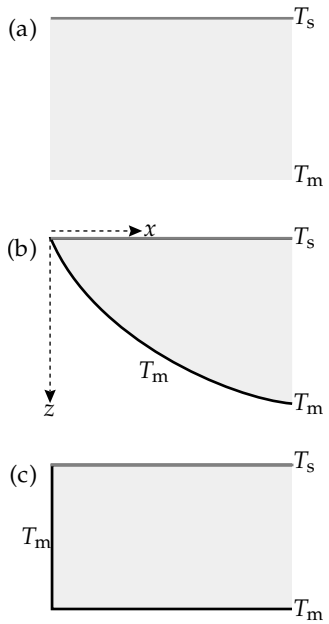

Fig. 4.32 (a) Semi-infinite half-space, (b) thermal boundary layer, and (c) plate models for the cooling of oceanic lithosphere. T_s and T_m are surface and mantle temperature, respectively.

process can be envisioned as transporting the column away from the ridge axis, during which conductive cooling takes place and the temperature distribution in the column changes.

This model allows us to compute the temperature distribution in the oceanic lithosphere. We need to compute the depth z at which a given temperature T is reached after time t, when the vertical column has moved at velocity v to a distance vt from the ridge. First, the desired temperature T is expressed as a fraction of the mantle temperature T_m. Using Eq. (4.57) and the appropriate table, the argument η_0 is found which gives an error function equal to (T/T_m). Setting the numerical value η_0 equal

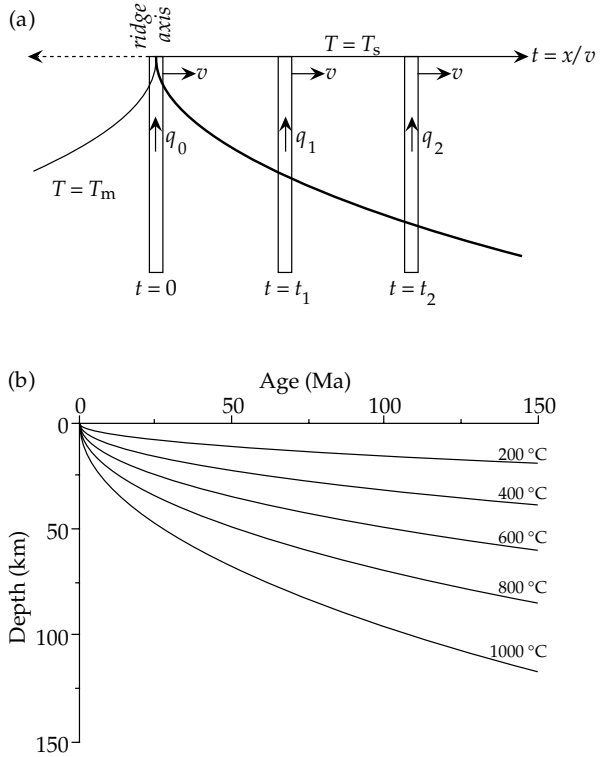

Fig. 4.33 Application of the infinite half-space model to explain the cooling of oceanic lithosphere: (a) vertical heat flow in narrow columns that move away from the ridge crest, and (b) predicted thermal structure in the cooling plate (after Turcotte and Schubert, 1982).

to $z/2\sqrt{\kappa t}$ gives the shape of the isotherm for the temperature T:

$$\frac{z}{2\sqrt{\kappa t}} = \eta_0$$

$$z = (2\eta_0\sqrt{\kappa})\sqrt{t} \tag{4.64}$$

The isotherms in the cooling lithosphere have a parabolic shape with respect to the time (or horizontal distance) axis, of which only the part for $z > 0$ is of interest. The surface heat flow for this model is given by Eq. (4.60), and so is inversely proportional to $\sqrt{t}$.

The half-space model has some unrealistic aspects. It predicts infinitely large heat flow at the ridge axis, and the initial mantle temperature T_m is approached asymptotically and is only reached at infinite depth. The distances between successive isotherms for equal increments in temperature get progressively larger. The near-surface layer in which the temperature changes are significant has been called a *thermal boundary layer*. Its base is defined arbitrarily as the depth at which the temperature reaches a chosen fraction of T_m. The layer can be regarded as a thermal model of the lithosphere (Fig. 4.32b). Instead of being defined mechanically as the depth where seismic shear waves are attenuated, the base of the lithosphere in the thermal model is an isotherm (Fig. 4.33b). The model predicts that the lithosphere becomes thicker with increasing age, as also inferred from seismic data, and the thickness is proportional to $\sqrt{t}$. As it cools and thickens, the lithosphere sinks deeper into the asthenosphere, so that the ocean depth increases away from a ridge axis. Together with Pratt-type, thermally influenced isostasy the half-space model of lithosphere cooling predicts a depth increase that is also proportional to $\sqrt{t}$ (Box 4.4).

Parker and Oldenburg (1973) proposed a modification of the boundary-layer model in which a solid lithosphere overlies a fluid asthenosphere. The base of the lithosphere is taken to be the solid–liquid phase boundary of the material. It is defined by the melting-point isotherm, and denotes a phase change. This is probably a closer representation of the real situation, although, by treating the asthenosphere as a fluid, it exaggerates the change in rheology. The temperature of the asthenosphere lies close to the solidus temperature, but its condition is only partially molten (perhaps about 5%).

The half-space and boundary-layer models fit the observed variations of heat flow and ocean depth with age for young lithosphere. For ages greater than about 70 Ma the ocean depths in particular are less than predicted by the $\sqrt{t}$ relationship (Fig. 4.28). This suggests that the source of heat from below the lithosphere may be shallower than in the half-space model at large ages. As an alternative to the half-space models the oceanic lithosphere has been modelled as a flat layer or plate of finite thickness, bounded above by cold sea-water and with a constant temperature on its lower surface. Far from a ridge axis this model brings hot mantle temperatures nearer to the surface than in the half-space model. Below the ridge axis the vertical edge of the new plate has the same high temperature of its lower surface (Fig. 4.32c), which results in heat being conducted horizontally through the plate. This is not a serious problem as long as the plate is much thinner than its horizontal extent away from a ridge. This condition is clearly met for the main lithospheric plates, which are several thousand kilometers across and only of the order of a hundred kilometers thick.

The plate model is not intended to model the vertical mechanical structure of the plate, but only to explain in a phenomenological way the typical age dependence of both ocean depth and heat flow. The plate thickness in the model is the asymptotic thermal thickness of old oceanic lithosphere and reflects the combined effects of temperature and rheology. Its horizontal isothermal base requires additional deep heat sources that prevent the lithosphere from cooling as a half-space at great ages. The model allows simple computation of the thermal cooling history. The best known version, proposed by Parsons and Sclater (1977), assumed a plate thickness of 125 km and a basal temperature of 1350 °C. At large distances from each spreading center it gives very good fits to both the observed heat flow (Fig. 4.27) and the ocean depth (Fig. 4.28). The most recent update, the GDH1 plate model, has a thickness of 95 km and a basal temperature of 1450 °C. It fits the observations even better.

Box 4.4: Variation of ocean depth with age

As oceanic lithosphere moves away from the ridge it cools, thickens, and becomes denser (Fig. 4.31). It sinks progressively into the underlying asthenosphere with time, so that the depth of an oceanic basin increases with age t of the oceanic lithosphere. A simple model accounts for the depth change w by assuming that the lithosphere and asthenosphere are in Pratt-type isostatic balance. The isostatic model is sometimes referred to as *thermal isostasy*.

Compare the composition of two vertical columns of unit cross-sectional area above a compensation level in the asthenosphere at depth D (Fig. B4.4). The column below R on the ridge axis consists of hot asthenosphere, of assumed constant density ρ_a and temperature T_a, and the depth d_r of sea-water (density ρ_w) above the ridge. The column below B over the adjacent ocean basin is a section through oceanic lithosphere. The density ρ_L of the lithosphere depends on its temperature T_L and the lithosphere thickness L increases with age t. The sea-water layer of depth d_r above the ridge is present in both columns, as is the thickness A of asthenosphere between the base of the lithosphere and the compensation depth. The isostatic balance is determined by equating the weights of w km of seawater and L km of lithosphere with the corresponding weight of $(w+L)$ km of asthenosphere:

$$(w+L)\rho_a g = w\rho_w g + g\int_0^L \rho_L \,\mathrm{d}z \tag{1}$$

$$w(\rho_a - \rho_w) = \int_0^L (\rho_L - \rho_a)\mathrm{d}z \tag{2}$$

Thermal isostasy assumes that the lithosphere changes density as it cools. The volume coefficient of expansion α is defined by Eq. (4.26) as

$$\alpha = \frac{1}{V}\frac{\mathrm{d}V}{\mathrm{d}T} = -\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}T} \tag{3}$$

where density $\rho = M/V$ and thus $\mathrm{d}V/V = -\mathrm{d}\rho/\rho$. Re-writing Eq. (3), we get

$$\alpha\rho_a(T_a - T_L) = (\rho_L - \rho_a) \tag{4}$$

The expression for the density difference between the lithosphere and asthenosphere is now substituted into Eq. (2):

$$w = \frac{\alpha\rho_a}{(\rho_a - \rho_w)}\int_0^L (T_a - T_L)\,\mathrm{d}z \tag{5}$$

The temperature of the lithosphere T_L is given by Eq. (4.57) with T_a instead of T_m. Substituting in Eq. (5), we find

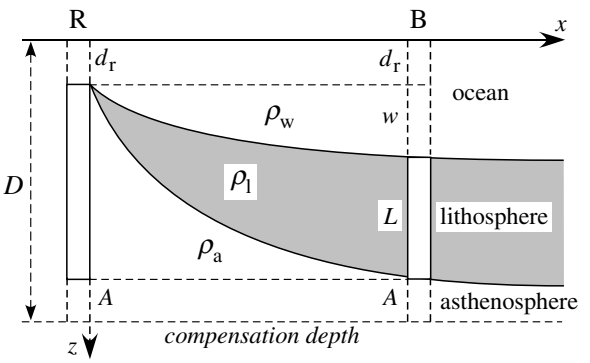

Fig. B4.4 Vertical section through oceanic lithosphere from a ridge to an adjacent ocean basin.

$$w = \frac{\alpha\rho_a}{(\rho_a - \rho_w)}\int_0^L (T_a - T_a\,\mathrm{erf}(\eta))\,\mathrm{d}z$$

$$= \frac{\alpha\rho_a T_a}{(\rho_a - \rho_w)}\int_0^L \mathrm{erfc}(\eta)\mathrm{d}z \quad \text{with} \quad \eta = \frac{z}{2\sqrt{\kappa t}} \tag{6}$$

The complementary error function, erfc(η), decreases almost to zero by $\eta = 2$, so the upper limit of integration can be changed from L to ∞ without causing significant error:

$$w = \frac{\alpha\rho_a T_a}{(\rho_a - \rho_w)}\int_0^\infty \mathrm{erfc}(\eta)\,\mathrm{d}z$$

$$= \frac{\alpha\rho_a T_a}{(\rho_a - \rho_w)} 2\sqrt{\kappa t}\int_0^\infty \mathrm{erfc}(\eta)\mathrm{d}\eta \tag{7}$$

From Box 4.3, Eqs. (8) and (9), we have $\int_0^\infty \mathrm{erfc}(\eta)\mathrm{d}\eta = 1/\sqrt{\pi}$. Thus,

$$w = \frac{2}{\sqrt{\pi}}\frac{\alpha\rho_a T_a}{(\rho_a - \rho_w)}\sqrt{\kappa t} \tag{8}$$

This is the amount by which the ocean deepens away from the ridge axis. The total depth of the ocean, taking into account the depth d_r at the ridge, is $d = (d_r + w)$. Optimum values for the parameters of the lithosphere in Eq. (8) are given in the Global Depth and Heat Flow model (GDH1) of Stein and Stein (1992): $\alpha = 3.1 \times 10^{-5}$ K^{-1}, $\rho_a = 3300$ kg m^{-3}, $\kappa = (k/c_p\rho_a) = 8.04 \times 10^{-7}$ m^2 s^{-1}. Assuming a mean depth of 2600 m over the ridge axis, a temperature $T_a = 1450$ C for the asthenosphere, and $\rho_w = 1030$ kg m^{-3} for the density of sea-water, the depth of the ocean over crust of age t (in Ma) is given by

$$d = (d_r + w) = 2600 + 370\sqrt{t} \tag{9}$$

This computed result is close to the depth–age relationship in Eq. (4.63) predicted by the GDH1 model for young lithosphere (age < 20 Ma).

4.2.8.4 Thermal structure of oceanic lithosphere

The plate models explain observed thermal data better than the boundary-layer models. The boundary-layer model is most appropriate near to a ridge axis, and agrees better with other geophysical data, which show that the lithosphere thickens with distance from a ridge. However, the plate model is needed at great distances to explain heat flux and ocean depths over old lithosphere. To reconcile these contrasting attributes a two-layered model of the lithosphere has been proposed (Fig. 4.34). The upper layer is rigid and has a mechanically defined lower boundary, above which heat transfer is by conduction. Below this level the increasing temperature causes a change in mechanical properties. The lower lithosphere is plastic enough to permit material movement, and so behaves like a viscous solid. The base of the upper layer is an isotherm, representing the temperature at which rigidity is lost. The base of the lower lithosphere is a thermally defined boundary, and is also an isotherm. Several suggestions have been made as to how this structure may approximate the plate model for old lithosphere. They include additional heat sources such as radiogenic heating, frictional heating as a result of shear at the base of the lithosphere, and reheating of old lithosphere due to the intrusion of mantle plumes at hotspots. It has also been postulated that, at lithosphere ages greater than about 70 Ma, small-scale convection currents in the lower lithosphere may augment thermal conduction. This would bolster the transfer of heat from the convecting asthenosphere into the lithosphere, effectively giving a thinner lithosphere than in the half-space models. Analysis of the dispersion of seismic surface waves indicates that there are differences in structure between continents and oceans down to about 200 km. This is compatible with the thermal model of a rigid mechanical layer underlain by a convecting thermal boundary layer extending to about 150–200 km.

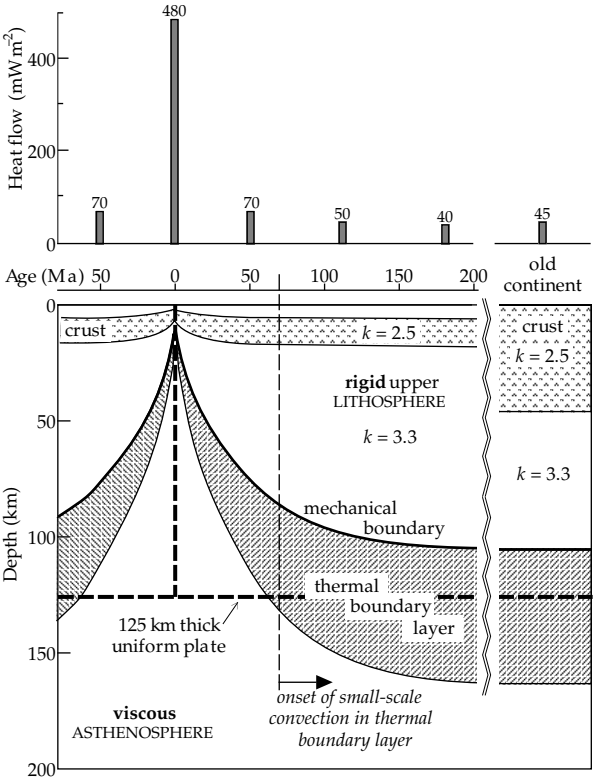

Fig. 4.34 Schematic diagram of lithospheric plate structure beneath oceans and continents. The dashed line indicates the approximation as a plate of constant thickness (based upon Parsons and McKenzie, 1978, and Sclater *et al.*, 1981).

4.2.8.5 Heat flow at subduction zones

The oceanic lithosphere is bent sharply downward beneath the overriding plate in a subduction zone. It extends as an inclined slab deep into the upper mantle, which it penetrates at a rate of a few centimeters per year. The old lithosphere is cold, having lost much of its original heat of formation at the ridge axis. By the time it reaches an ocean trench the isotherms in the plate are far apart and the temperature gradient is small. The separation of the isotherms is increased by the downward bending of the plate. Since the heat flow is proportional to the temperature gradient, very low heat-flow values ($\approx 35\,\text{mW m}^{-2}$) are measured in oceanic trenches. After bending downward the plate is subducted to great depths, subjecting it to increases in pressure and temperature. Heat is conducted into the plate from the adjacent mantle. This process is so slow that the interior of the subducting slab remains colder than its environment (Fig. 4.35). A temperature of 800 °C is normally reached at about 70 km depth in the oceanic plate but in the descending slab this temperature exists to deeper than 500 km. Above this depth the coldest part of the slab has a horizontal temperature deficit of 800–1000 K.

Heat conducted from the mantle is not the only heat source that must be taken into account in modelling the thermal structure of the subducting slab. An important additional source is the frictional heating that results from shear deformation at the surfaces of the slab where it is in contact with the mantle. In the upper part of a subduction zone the shear heating melts the basaltic layer of the oceanic lithosphere and forms a layer of eclogite in the top of the slab. The high density of the eclogite causes a positive gravity anomaly (see Fig. 2.62), and adds to the forces propelling the slab downward. The phase transition in which the open structure of olivine-type minerals converts to a denser spinel-type structure normally takes place at a depth of 400 km. The phase transition depends on temperature and pressure. Laboratory experiments indicate that it takes place at lower pressure at low temperature than at high temperature. Consequently it occurs at shallower depths within the cold plate than in the adjacent mantle. As a result the transition depth is deflected upward by about 100 km. The transition is exothermic and the latent heat given out in the transition is an additional heat source that contributes to the thermal structure of the subduction

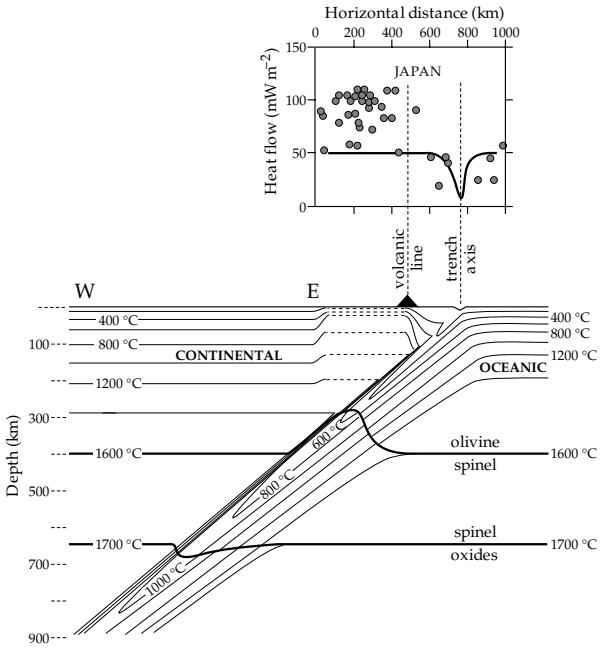

Fig. 4.35 *Bottom*: the thermal structure of a subduction zone and back-arc region (the model of Schubert *et al.*, 1975, inverted horizontally), showing the possible isotherms in the cold subducting plate and the thermal effects of the olivine–spinel and spinel–oxide phase changes. *Top*: comparison of heat-flow measurements across the Japanese trench with the theoretical heat flow (solid curve) computed by Toksöz *et al.* (1971).

zone. The transition also results in a density increase, which adds to the forces driving the plate downward.

The deeper transition at 670 km is less well understood. High temperature apparently causes it to take place at higher pressure, and so the depth of occurrence is deflected downward inside the subducting slab. It is uncertain whether the transition is endothermic, absorbing heat from the environment, or exothermic as assumed in the model in Fig. 4.35. An endothermic phase change has the effect of reducing the density, and acts against the other downward forces on the slab.

Although other, slightly different models have been derived for the temperature distribution in the descending slab, they all have in common the downward deflection of isotherms in the cold descending slab. The heat flow can be computed for a given thermal model. When compared with the observed heat flow on a profile across the subduction zone, the models fail to explain adequately the high heat flow observed on the overriding plate (Fig. 4.35). Volcanic activity is partly responsible, fed by magmas produced by partial melting of oceanic crustal material in the descending slab and of the upper mantle in the overriding plate. Shallow melting is promoted by water from the subducting plate and generates basaltic magma; deeper melting involves less water and results in andesitic magma. When the overriding plate is continental, volcanic chains form along the continental margin parallel to the deep oceanic trench. The volcanicity is typified by the eruption of both basaltic and andesitic lavas. The lavas are more felsic than those formed when two oceanic plates collide, which may imply that they include melted material from the upper mantle of the overriding continental plate.

When two oceanic plates converge, a volcanic arc is formed on the overriding plate. Behind the arc, high heat flow on the overriding plate is related to back-arc spreading, in which new oceanic crust is generated by the intrusion of basaltic magma from partial melting in the upper mantle. This form of sea-floor spreading produces a marginal basin behind the island arc. The intrusion of magma is not confined to a single location, as at a ridge axis, but is spread diffusely in the basin. Consequently, the stripes of lineated oceanic magnetic anomalies characteristic of sea-floor spreading at ridge systems are missing or at best weakly defined in a marginal basin.

4.2.9 Mantle convection

It has gradually become accepted that thermally driven convection takes place in the mantle and that it is probably the most important mechanism in geodynamic processes. There are several reasons for these conclusions. The evidence summarized in Section 2.8 demonstrates that the mantle has a viscoelastic rheology. The passage of seismic compressional and shear waves through the mantle attest that it reacts as a solid to abrupt stress changes. Yet, observations of post-glacial isostatic uplift and long-term movements of the rotation axis indicate that the mantle is capable of viscous flow when stressed over long time intervals. The surmised temperature distribution in the mantle implies that, although conduction is mainly responsible for heat transfer in the lithosphere, convection is the predominant process deeper in the mantle, involving mass transfer by sub-solidus creep. Applying the theory of thermal convection to the mantle and using the best available estimates of physical parameters indicates that robust convection must be taking place.

4.2.9.1 Thermal convection

The conditions for convection to occur (see Section 4.2.4.2) reflect a balance between causal forces due to thermal expansion and resistive effects due to viscosity and thermal diffusivity. When a fluid is heated, thermal expansion gives rise to an upward buoyancy force. This produces instability, which is partly counteracted by diffusion of heat into the surrounding fluid by thermal conduction. As soon as a volume of the fluid starts to rise in response to the buoyancy force its motion is resisted by viscous forces. The effects are familiar to anyone who has heated a pan of thick soup or porridge. If the pan is heated too rapidly, or the instructions to "stir constantly" are ignored, the soup may stick to the bottom of the pan and become charred. This happens because the viscosity of the fluid is initially too large to allow convection. Despite the large temperature gradient between the hot bottom of the pan and the cool surface of the liquid, conduction is unable to

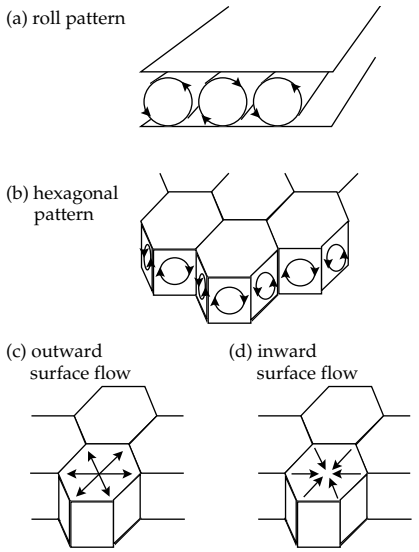

Fig. 4.36 Some patterns of steady convection in a plane layer heated from below. (a) Convection rolls, (b) vertical flow in hexagonal patterns, for which the surface flow may be (c) outward away from or (d) inward toward the center of the cell (after Busse, 1989).

transport heat away from the bottom of the pan fast enough to avoid charring. When heat flow by conduction reaches a critical limit, convection can begin.

The onset of convection in a fluid layer heated from beneath was first described in 1900 by H. Bénard on the basis of laboratory experiments. He noted that a hexagonal pattern of cells forms on the surface of the layer (Fig. 4.36). Hot fluid rises to the surface in the middle of each cell; at the surface it spreads out and cools. Adjoining cells come in contact at narrow margins, where the cooled fluid sinks back into the layer. Each cell has a rectangular cross-section in the vertical plane. A satisfactory theory of Bénard's observations was derived in 1916 by Lord Rayleigh. Although it applies to an ideal scenario (a horizontal layer with stress-free upper and lower boundaries, heated from below, and with a constant temperature on the upper surface) the theory permits approximate estimates for more complex convection in the spherical Earth.

The flow of a viscous fluid is governed by the Navier–Stokes equation, one of the most important equations in geophysics. It describes the conservation of momentum in the fluid, which in its simplest form means balancing several terms that express the driving forces exerted by pressure gradient and buoyancy against the viscous and inertial forces that resist motion. The ratio of the other forces to the inertial forces is expressed by the dimensionless Prandtl number, Pr, defined as

$$\mathrm{Pr} = \frac{\nu}{\kappa} \tag{4.65}$$

where ν is the kinematic viscosity, and κ is the thermal diffusivity. In the mantle $\nu \approx 10^{18}$ m^2 s^{-1} and $\kappa \approx 10^{-6}$ m^2 s^{-1}, so that $\mathrm{Pr} \approx 10^{24}$. The virtually infinite Prandtl number means that inertial forces are insignificant. Hence, mantle convection depends only on the conditions of pressure, temperature and viscosity.

Thermal diffusivity and viscosity act as stabilizing influences in a heated fluid. If heating is slow enough, the temperature gradient adjusts to transfer the heat by conduction, remaining close to the adiabatic gradient. Convection becomes possible when the real temperature gradient exceeds the adiabatic gradient; the difference θ is called the *superadiabatic gradient*. The excess heat expands the fluid, causing the buoyancy force. When this becomes larger than the viscous resistance, convection ensues. The ratio of the competing forces is embodied in the *Rayleigh number* (Eq. (4.42)). The Rayleigh number (Ra_T) for convection due to the superadiabatic temperature gradient in a fluid layer of thickness D is

$$\mathrm{Ra}_T = \frac{g\alpha\theta}{\kappa\nu} D^4 \tag{4.66}$$

where g is gravity and α is the coefficient of thermal expansion.

The superadiabatic gradient is not the only source of power for convection. Although radioactive heat generation in mantle materials is small (see Section 4.2.5.1), it can still contribute to convection. If Q is the radiogenic heat production in a layer of thickness D, we can invoke Eq. (4.39) and Eq. (4.45) and replace θ in the above equation by (QD/k), where k is the thermal conductivity. This allows us to define a second Rayleigh number (Ra_Q) for convection driven by radiogenic heat:

$$\mathrm{Ra}_Q = \frac{g\alpha Q}{k\kappa\nu} D^5 \tag{4.67}$$

Convection is initiated when the Rayleigh number exceeds a critical value, Ra_c, which is dependent on the geometry of the flow and the boundary conditions on the upper and lower surfaces. In Rayleigh–Bénard convection the top and bottom of the horizontal layer are stress free; the critical Rayleigh number is $\mathrm{Ra}_c = 658$. If the top and bottom of the layer are rigid boundaries at which the horizontal velocity vanishes, $\mathrm{Ra}_c = 1708$. Table 4.8 shows computed Rayleigh numbers Ra_T for convection driven by the superadiabatic temperature gradient for viscous flow in the upper, lower and whole mantle, assuming representative values from the literature for the parameters in Eq. (4.66). Reasonable estimates of the radiogenic heat produced in the mantle give even larger values for Ra_Q.

4.2.9.2 Convection at high Rayleigh numbers

The computed Rayleigh numbers greatly exceed the critical values for convection throughout the entire mantle or in separate layers. Thus, each region of the sub-lithospheric mantle is capable of convection. The Rayleigh number for whole-mantle convection is so much larger than the critical value Ra_c that vigorous mantle convection must be expected. This does not imply rapid flow in normal terms. The speed of flow in the mantle is usually assumed to be of the same order as the rate of motion of tectonic plates, about 5–10 cm yr^{-1} on average. As long as the flow rate v is

Table 4.8 *Some physical parameters for mantle convection models (mostly from Jarvis and Peltier, 1989)*

The critical Rayleigh numbers (Ra_c) for the onset of convection in each part of the mantle are calculated assuming a superadiabatic temperature gradient $\theta = 0.1$ K km^{-1} and a mean gravity $g = 10$ m s^{-2}. Lower mantle parameters are interpolated from the upper- and whole-mantle values.

Physical parameter	Units	Upper mantle (70−670 km)	Lower mantle (670−2890 km)	Whole mantle (70−2890 km)
Layer thickness (H)	km	600	2220	2820
Expansion coefficient (α)	K^{-1}	2×10^{-5}	1.0×10^{-5}	1.4×10^{-5}
Density (ρ)	kg m^{-3}	3700	5500	4700
Specific heat (c_p)	J kg^{-1} K^{-1}	1260	1260	1260
Thermal conductivity (k)	W m^{-1} K^{-1}	6.7	20	15
Thermal diffusivity (κ)	m^2 s^{-1}	1.4×10^{-6}	3×10^{-6}	2.5×10^{-6}
Dynamic viscosity (η)	kg m^{-1} s^{-1}	1×10^{21}	2.5×10^{21}	2×10^{21}
Kinematic viscosity (ν)	m^2 s^{-1}	2.7×10^{17}	4.5×10^{17}	4.3×10^{17}
Rayleigh number (Ra_T)	—	7000	180,000	820,000

low, adjacent lamina of the fluid move past each other under the conditions for Newtonian viscosity (Section 2.8.2). At faster flow rates this condition breaks down, and the flow becomes turbulent. The conditions favoring turbulence are high momentum (ρv) and large scale D of the flow, whereas it is inhibited by high viscosity η. These factors are contained in the Reynolds number, Re, defined as

$$\mathrm{Re} = \frac{\rho v D}{\eta} \tag{4.68}$$

Reasonable values for the mantle are $\rho = 5000$ kg m^{-3}, $D = 2900$ km $= 2.9 \times 10^6$ m, $v = 5$ cm yr$^{-1} = 1.5 \times 10^{-9}$ m s^{-1}, and $\eta = 1.5 \times 10^{21}$ Pa s. The Reynolds number is found to be $\mathrm{Re} = 1.5 \times 10^{-20}$, which is so small that turbulence is negligible. Similar results are found by considering the upper or lower mantle alone. Clearly, although mantle convection involves high Rayleigh numbers (implying vigorous convection on a geological timescale), it takes place by laminar flow.

The effect of convection is to replace conduction as the principal mechanism of heat transfer. A measure of the relative effectiveness of the two processes of heat transfer is the Nusselt number, Nu. This is defined as the ratio of the heat transport in the presence of convection to the heat transport without convection. In the absence of radiogenic heat sources, the heat transport with convection is determined by the Rayleigh number Ra_T, while the non-convective heat transport is expressed by the critical Rayleigh number Ra_c. The Nusselt number depends on the ratio of these two numbers and can be written

$$\mathrm{Nu} = \beta \left(\frac{\mathrm{Ra}_T}{\mathrm{Ra}_c} \right)^S \tag{4.69}$$

where the coefficient β and the exponent S are functions of the aspect ratio of the convection cells. Mathematical evaluation of the problem of Rayleigh–Bénard convection with stress-free upper and lower boundaries gives $\beta \approx 1$ and $S = 1/3$, and, since in this case $Ra_c \approx 10^3$, the Nusselt number has the simpler form

Table 4.9 *Approximate aspect ratios of some mantle convection cells, estimated from the horizontal dimensions of the overlying lithospheric plates (after Turcotte and Schubert, 1982, Table 7.5)*

Plate	Upper-mantle convection	Whole-mantle convection
Pacific	14	3.3
North American	11	2.6
South American	11	2.6
Indian	8	2.1
Nazca	6	1.6

$$\mathrm{Nu} \approx 0.1 (\mathrm{Ra}_T)^{1/3} \tag{4.70}$$

Using the estimated values of Ra_T in Table 4.8 gives Nusselt numbers of 19 for layered convection in the upper mantle and 97 for whole-mantle convection. Hence, heat transfer by convection is dominant in the mantle.

Once convection has been initiated the boundary conditions determine the shapes of the convection cells. In Rayleigh–Bénard convection the aspect ratio of a cell – the ratio of its horizontal dimension to its vertical one – is $2^{1/2} = 1.41$; when the layer has rigid boundaries the cell aspect ratio is 1.01. Hence, the horizontal extent of a convection cell is comparable with the layer thickness. This has implications for convection in the mantle. If we assume that the scale of mantle convection is represented by the pattern of the plate boundaries (Fig. 1.11), we can estimate the horizontal dimensions of the convection cells. It is evident that they must have very different sizes. The ridge-to-trench horizontal distances across major plates are in the range 2000–10,000 km, with an average of about 5000 km. This is larger than the maximum thickness of the convecting layer, whether we assume convection to be restricted to the upper mantle or to occupy the entire 2900 km thickness of the sub-lithospheric mantle. Thus, if convection is uniform through the whole mantle, the aspect ratios of at least some cells must be much larger than unity (Table 4.9). The

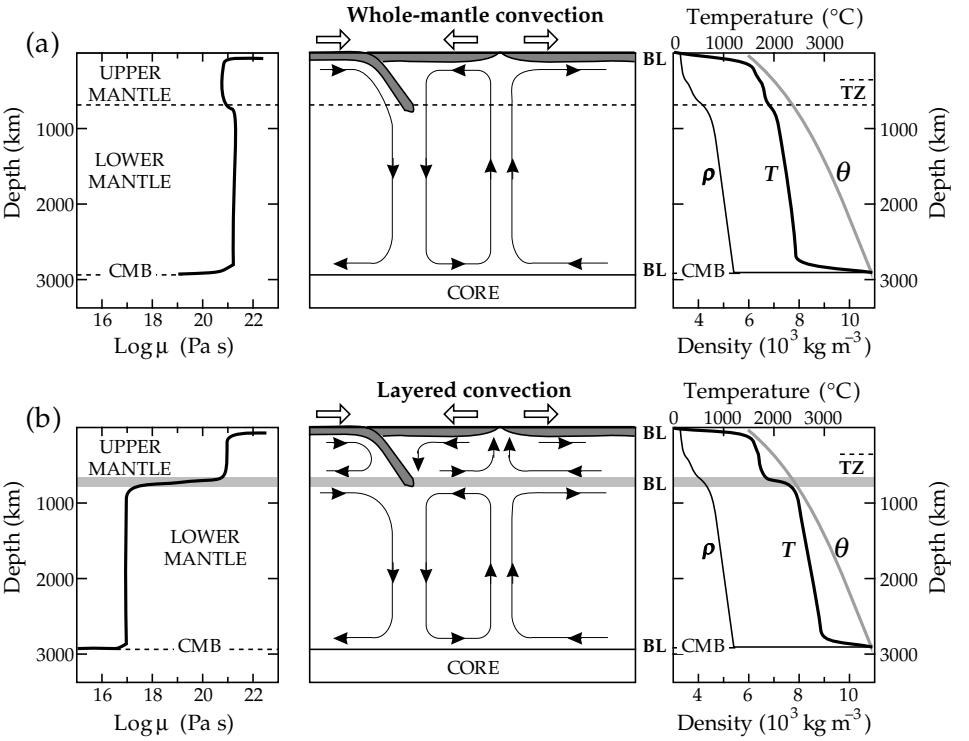

Fig. 4.37 Possible convection flow pattern (*center*) and profiles of viscosity μ (*left*), and density ρ, temperature T and solidus temperature θ (*right*) for (a) whole-mantle convection and (b) layered mantle convection. TZ is the upper-mantle transition zone, BL are boundary layers, CMB is the core–mantle boundary (based upon Peltier *et al.*, 1989).

reason for this is the rigidity of the cold upper boundary formed by the lithosphere, which inhibits the breakup of the fluid flow into cells with smaller horizontal extents.

4.2.9.3 Models of mantle convection

The feasibility of mantle convection is accepted but there is still some doubt as to the form it takes. This is in part due to uncertainty as to the role played by the seismic discontinuities at 400 km and 670 km depth, which bound the upper-mantle transition zone. The discontinuities are not sharp, and are understood to represent mineral phase changes rather than compositional differences (as, for example, the crust–mantle and core–mantle boundaries). The upper discontinuity marks the olivine–spinel phase change, the lower one represents the phase change from spinel to perovskite structure (Section 3.7.5.2), with accompanying changes in density and elastic parameters. In principle, mass can be carried by convection currents across these discontinuities. The 670 km discontinuity is close to the maximum depth of seismicity in subduction zones, and may be where the subducting plate is absorbed into the mantle.

There are two main models of mantle convection, each with an interface at the 670 km seismic discontinuity. An important change in viscosity occurs at this level. In *whole-mantle convection* (Fig. 4.37a) the viscosity doubles from the upper mantle to the lower mantle (see Table 4.8) and there is a net flow of material across the boundary. In this model, convection ensures that the entire mantle is well mixed mechanically, and the phase changes at 400 and 670 km have only a small effect on the temperature gradient. This model agrees with much of the available evidence.

The alternative *layered convection* model has distinct convecting layers in the upper and lower mantle (Fig. 4.37b). There are two ways in which this can take place. The upper and lower convection patterns in a vertical section may represent circulations in the same sense (e.g., both clockwise or both counterclockwise) or in opposite senses (e.g., one clockwise and the other counterclockwise). In each case the radial velocity is zero at 670 km depth and there is no mass transfer across the discontinuity; the material in each flow pattern spreads out along the boundary. However, the models imply different types of coupling between the layers. Opposite senses of circulation in the layers would cause little or no shear between the tangential flows at the boundary, resulting in mechanical coupling between the layers. Cold material sinking in the upper mantle would overlie hot material rising in the lower mantle. However, if the layered flow patterns have the same sense of circulation (as in Fig. 4.37b), hot material rising in the upper mantle overlies hot material rising in the lower mantle, so that the flow regimes are coupled thermally. This model has a strong velocity shear across the 670 km discontinuity, which requires a large and abrupt change in viscosity at this depth; viscosity in the lower mantle would need to be at least two orders of magnitude smaller than in the upper mantle. Estimates of mantle viscosity (Section 2.8.6) indicate the opposite: viscosity is higher in the lower mantle than in the upper mantle.

A model of layered convection assumes that there is no mass transfer across the discontinuity. The upper and lower mantles are well mixed individually, but the separation of the flow patterns at the discontinuity means that they may have distinct chemical compositions. Because

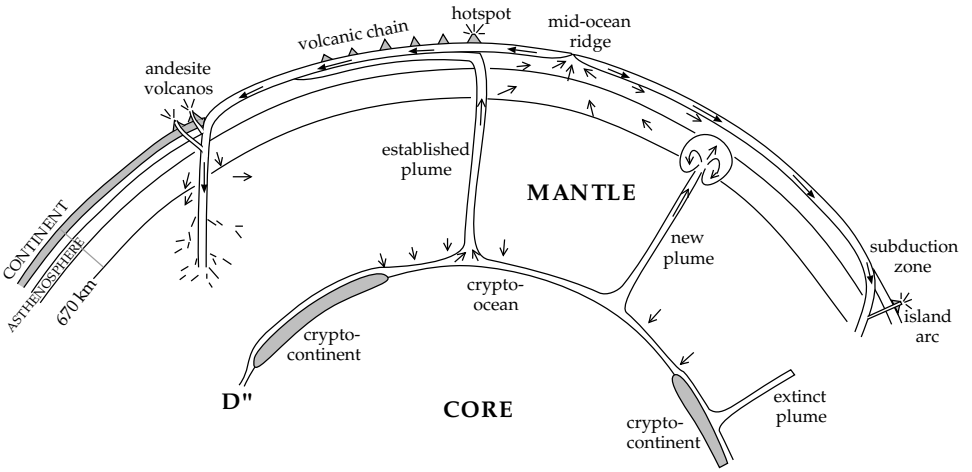

Fig. 4.38 An idealized cross-section through the mantle, showing convective flow and the relationship of mantle plumes to the D″-layer (after Stacey, 1992).

there is no convective flow across it, heat can only cross the boundary by conduction. The 670 km discontinuity therefore acts as a thermal boundary, with a large temperature change of perhaps 500–1000 K across it. Thus, the temperature profile in the lower mantle, although maintained adiabatic by the convection, would be 500–1000 K higher than in whole-mantle convection. This would result in a smaller temperature change across the core–mantle boundary, a less-steep temperature gradient in the D″-layer, and so a lower heat flux from the core. The long-term rate of cooling of the Earth would thereby be reduced.

The problem of understanding mantle convection is complicated by the non-uniform structure and rheology of the mantle. As yet, there is no complete picture of how the various factors that influence convection act together. The convection pattern depends strongly on what happens physically and thermodynamically at the 670 km discontinuity. This can only be inferred indirectly. Our understanding of the discontinuity is incomplete, but it is essential to resolving the real pattern of mantle convection.

4.2.9.4 Mantle plumes

The viscosity in the upper mantle is inferred from post-glacial rebound studies to be around 10^{21} Pa s, but lower-mantle viscosity is less well known. The sub-solidus creep in the mantle implies a temperature-dependent viscosity, which allows thermal boundary layers at the top and bottom of the mantle to influence the patterns of convective flow.

The lithosphere constitutes an upper, cold boundary layer. It accretes at high temperature at spreading ridges, where upwelling magma from the mantle reaches the surface. The eruptive lavas issue from magma chambers beneath ridge crests, in which magma from the deeper mantle undergoes differentiation. As part of the plate tectonic cycle the lithosphere rapidly cools and hardens as it spreads away from the ridge. Its high viscosity (i.e., rigidity) inhibits internal convection, but at subduction zones the plate (by now old) flexes downward and carries cold material into the underlying mantle, altering its thermal balance. Seismic tomography (Section 3.7.6) has revealed broad regions of raised seismic velocity in the deep mantle below subduction zones, giving rise to the surmise that the material in the cold subducted plate eventually sinks to the bottom of the mantle. The material must eventually take part in a broad-scale return flow, completing the convective cycle, but how this takes place is not clear.

The core–mantle boundary (CMB) at 2890 km depth constitutes a lower, hot boundary layer. The D″-layer at the base of the mantle (Section 3.7.5.3) is characterized by reductions in seismic velocities between about 2740 km depth and the CMB. It evidently has different physical properties than the mantle above it, and appears to play an important thermodynamic role. The heat flux from the core to the mantle diminishes the rigidity of the layer, thus reducing the seismic velocities. The viscosity of the hot thin D″-layer is presumed to be much lower than that of the overlying mantle. The topography of the CMB has been explored by seismic waves reflected from the core or passing it at grazing incidence. The thickness of the D″ layer appears to be uneven, and has been interpreted by analogy to the crust. Thick segments have been designated as crypto-continents and thinner regions as crypto-oceans (Fig. 4.38).

The low-viscosity material in the D″-layer is thought to supply relatively fast-flowing narrow mantle plumes. This name is given to vertical features, thin in cross-section, that facilitate the upwelling of low-viscosity hot magma through the more viscous mantle. A new plume melts its way to the surface behind a larger head. Some plumes may not reach the surface but intrude their material into the asthenosphere or lower lithosphere. Other mature plumes may penetrate the entire mantle and reach the surface, where they are evident as places of persistent volcanism, high regional topography and local high heat flow, called *hotspots*. These areas of anomalous volcanism are found in the oceans and on the continents, within plates and on plate margins. The plumes that feed them are thought to remain fixed in position for long periods of

time, and so the hotspots are anchored to the mantle below the lithosphere. As a result they have important consequences for studies of plate tectonic motions.

4.3 GEOELECTRICITY

4.3.1 Introduction

Electric charge – together with mass, length and time – is a fundamental property of nature. The name *electric* derives from the Greek word for amber ("elektron"), the naturally occurring fossilized resin of coniferous trees that has been used since antiquity in the making of jewelry. The Greek philosopher Thales of Miletus (ca. 600 BC) is credited with first reporting the power of amber, when rubbed with a cloth, to attract light objects. The ancient sages could not understand this behavior in terms of their everyday world, and so, together with the power of *magnetism* possessed by natural lodestone (see Section 5.1.1), electricity remained a wonderful but unknown phenomenon for more than two millennia. In 1600 AD the English physician William Gilbert summarized previous investigations and extant knowledge in the first systematic study of these phenomena.

In the following century it was established that there were two types of electric charge, now referred to as positive and negative. Objects that carried like types of charge were observed to repel each other, and those that carried opposite types were attracted to each other. In 1752 the American statesman, diplomat and scientist Benjamin Franklin performed a celebrated experiment; by flying a kite during a thunderstorm, he established that lightning is an electrical phenomenon. Having survived this risky endeavor Franklin developed the far-sighted theory that electricity consisted of an omnipresent fluid, and that the different types of charge represented surplus and scarcity of this fluid. This view strikingly resembles modern theory, in which the "fluid" consists of electrons.

The laws of electrostatic attraction and repulsion were established in 1785 as a result of careful experiments by a French scientist, Charles Augustin de Coulomb (1736–1806), who also established the laws of magnetostatic force (Section 5.1.3). Coulomb invented a sensitive torsion balance, with which he could measure accurately the force between electrically charged spheres. His results represent the culmination of knowledge of electrostatic phenomena.

The eighteenth century concept of electricity as a fluid finds further expression in electrical nomenclature. Electricity is said to flow between charged objects when they are brought in contact, and the rate of flow is called an electric current. The study of the properties and effects of electric currents became possible around 1800, when an Italian physicist, Alessandro Volta (later elevated by Napoleon to the rank of Count), invented a primitive electric battery, called a voltaic pile, in which electricity was produced by chemical action. The relationship between the electric current in a conductor and the voltage of the battery was established in 1827 by Georg Ohm, a German physicist. The magnetic effects produced by electric currents were established in the early nineteenth century by Oersted, Ampère, Faraday and Lenz. Their contributions are discussed in more detail in a later chapter (Section 5.1.3) on the physical origins of magnetism.

4.3.2 Electrical principles

Coulomb established that the force of attraction or repulsion between two charged spheres was proportional to the product of the individual electric charges and inversely proportional to the square of the distance between the centers of the spheres. His law can be written as the following equation:

$$F = K\frac{Q_1 Q_2}{r^2} \tag{4.71}$$

where Q_1 and Q_2 are the electric charges, r is their separation and K is a constant. This inverse-square law strongly resembles the law of universal gravitation (Eq. (2.2)), formulated by Newton more than a century before Coulomb's law. However, in gravitation the force is always attractive, whereas in electricity it may be attractive or repulsive, depending on the nature of the charges. In the law of gravitation the units of mass, distance and force are already defined, so that the gravitational constant is predetermined; only its numerical value needed to be measured. In Coulomb's law, F and r are defined from mechanics (as the newton and meter, respectively), but the units of Q and K are undefined. The value of K was originally set equal to unity, thereby defining the unit of electric charge. This definition led to unfortunate complications when the magnetic effects of electric currents were analyzed. The alternative is to define independently the unit of charge, thereby fixing the meaning of the constant K.

The unit of charge is the coulomb (C), defined as the amount of charge that passes a point in an electrical circuit when an electric current of one ampère (A) flows for one second (i.e., $1\,\mathrm{C} = 1\,\mathrm{A\ s}$). In turn, the ampère is defined from the magnetic effects of a current (see Section 5.2.4). When a current flows in the same direction through two parallel long straight conductors, magnetic fields are produced around the conductors, which cause them to attract each other. If the current flows through the conductors in opposite directions they repel each other. The ampère is defined as the current that produces a force of 2×10^{-7} N per meter of length between infinitely long thin conductors that are one meter apart in vacuum. Thus, the unit of charge is defined precisely, if rather indirectly. In the Système Internationale (SI) units K is written as $(4\pi\,\varepsilon_0)^{-1}$, so that Coulomb's law becomes

$$F = \frac{1}{4\pi\varepsilon_0}\frac{Q_1 Q_2}{r^2} \tag{4.72}$$