INSTITUTE OF PHYSICS PUBLISHING REPORTS ON PROGRESS IN PHYSICS

Rep. Prog. Phys. **67** (2004) 1429–1496 PII: S0034-4885(04)25227-7

The physics of earthquakes

Hiroo Kanamori[1] **and Emily E Brodsky**[2]

[1] Seismological Laboratory, California Institute of Technology, Pasadena, CA 91125, USA
[2] Department of Earth & Space Sciences, University of California, Los Angeles, Los Angeles, CA 90095, USA

Received 21 January 2004, in final form 20 April 2004
Published 12 July 2004
Online at stacks.iop.org/RoPP/67/1429
doi:10.1088/0034-4885/67/8/R03

Abstract

Earthquakes occur as a result of global plate motion. However, this simple picture is far from complete. Some plate boundaries glide past each other smoothly, while others are punctuated by catastrophic failures. Some earthquakes stop after only a few hundred metres while others continue rupturing for a thousand kilometres. Earthquakes are sometimes triggered by other large earthquakes thousands of kilometres away. We address these questions by dissecting the observable phenomena and separating out the quantifiable features for comparison across events. We begin with a discussion of stress in the crust followed by an overview of earthquake phenomenology, focusing on the parameters that are readily measured by current seismic techniques. We briefly discuss how these parameters are related to the amplitude and frequencies of the elastic waves measured by seismometers as well as direct geodetic measurements of the Earth's deformation. We then review the major processes thought to be active during the rupture and discuss their relation to the observable parameters. We then take a longer range view by discussing how earthquakes interact as a complex system. Finally, we combine subjects to approach the key issue of earthquake initiation. This concluding discussion will require using the processes introduced in the study of rupture as well as some novel mechanisms. As our observational database improves, our computational ability accelerates and our laboratories become more refined, the next few decades promise to bring more insights on earthquakes and perhaps some answers.

(Some figures in this article are in colour only in the electronic version)

Contents

List of frequently used symbols

A	constant in the rate- and state-friction law
a	half-length of Mode III crack
α	P-wave speed
B	constant in the rate- and state-friction law
b	slope of the earthquake magnitude–frequency relationship
β	S-wave speed
χ	probability of earthquake occurrence
D	fault slip offset
$\bar{D}$	average offset
D_0	critical slip of a crack
D_c	critical displacement in the slip-weakening models
δ	slip on a frictional surface
$\dot{\delta}$	slip speed
E	elastic modulus
E_R	radiated seismic energy
E_G	fracture energy of the earthquake
E_H	thermal energy (frictional energy loss) of the earthquake
$\tilde{e}$	scaled energy (the ratio of radiated seismic energy to seismic moment)
G	dynamic energy release rate (dynamic crack extension force)
G^*	static energy release rate (static crack extension force, specific fracture energy)
G_c^*	critical specific fracture energy
γ	surface energy
η	viscosity, seismic efficiency
η_R	radiation efficiency
K	stress intensity factor
K_c	fracture toughness (critical stress intensity factor)
k	stiffness of spring, permeability
k_f	stiffness of the fault
$\tilde{L}$	length scale of the fault
$\tilde{L}_n$	nucleation length
l_0	crack breakdown length
M_0	seismic moment
M_w	earthquake magnitude (moment magnitude)
μ	rigidity (shear modulus) or coefficient of friction
μ_s	coefficient of static friction
μ_k	coefficient of kinetic friction
p	pore pressure, power of the stress–corrosion relation, power of Omori's Law
Q	heat
R	seismicity rate
r_0	background seismicity rate
ρ	density
S	fault area
σ_0	initial stress

σ_1	final stress (sections 3 to 6)
σ_f	frictional stress
$\Delta\sigma_s$	static stress drop ($\sigma_0 - \sigma_1$)
σ_{ij}	stress tensor
$(\sigma_1, \sigma_2, \sigma_3)$	principal stresses (section 2)
σ_Y	yield stress
σ_n	normal stress
τ	shear stress, source duration
$\bar{\tau}$	average source duration
$\dot{\tau}$	stress rate
θ	state variable in rate- and state-dependent friction; angle between the fault and the maximum compressional stress
u_i	displacement vector
V	rupture speed
W_0	initial (before an earthquake) potential energy of the Earth
W_1	final (after an earthquake) potential energy of the Earth
ΔW	change in the potential energy
ΔW_0	change in the potential energy minus frictional energy
w	width of the fault slip zone

1. Introduction

Why do earthquakes happen? This age-old question was solved at one level by the plate tectonics revolution in the 1960s. Large, nearly rigid plates of the Earth slide past each other. Earthquakes accommodate the motion (figure 1). However, this simple answer is far from complete. Some plate boundaries glide past each other smoothly, while others are punctuated by catastrophic failures. Why is so little motion accommodated by anything in between these two extremes? Why do some earthquakes stop after only a few hundred metres while others continue rupturing for a thousand kilometres? How do nearby earthquakes interact? Why are earthquakes sometimes triggered by other large earthquakes thousands of kilometres away?

Earthquake physicists have attempted to answer these questions by dissecting observable phenomena and separating out the quantifiable features for comparison across events. We begin this review with a discussion of stress in the crust followed by an overview of earthquake phenomenology, focusing on the parameters that are readily measured by current seismic techniques. We briefly discuss how these parameters are related to the amplitude and frequencies of the elastic waves measured by seismometers as well as direct geodetic measurements of the Earth's deformation. We then review the major processes thought to be active during rupture and discuss their relationship to the observable parameters. We then take a longer range view by discussing how earthquakes interact as a complex system. Finally, we combine subjects to approach the key issue of earthquake initiation. This concluding discussion will require using the processes introduced in the study of rupture, as well as some novel mechanisms.

In this introductory review for non-specialists, we gloss over many exciting and important advances in recent years ranging from the discovery of periodic slow slip events (Dragert *et al* 2001) to the elucidation of fault structure revealed by new accurate location techniques (Rubin *et al* 1999). Many of these recent advances are made possible by new technology

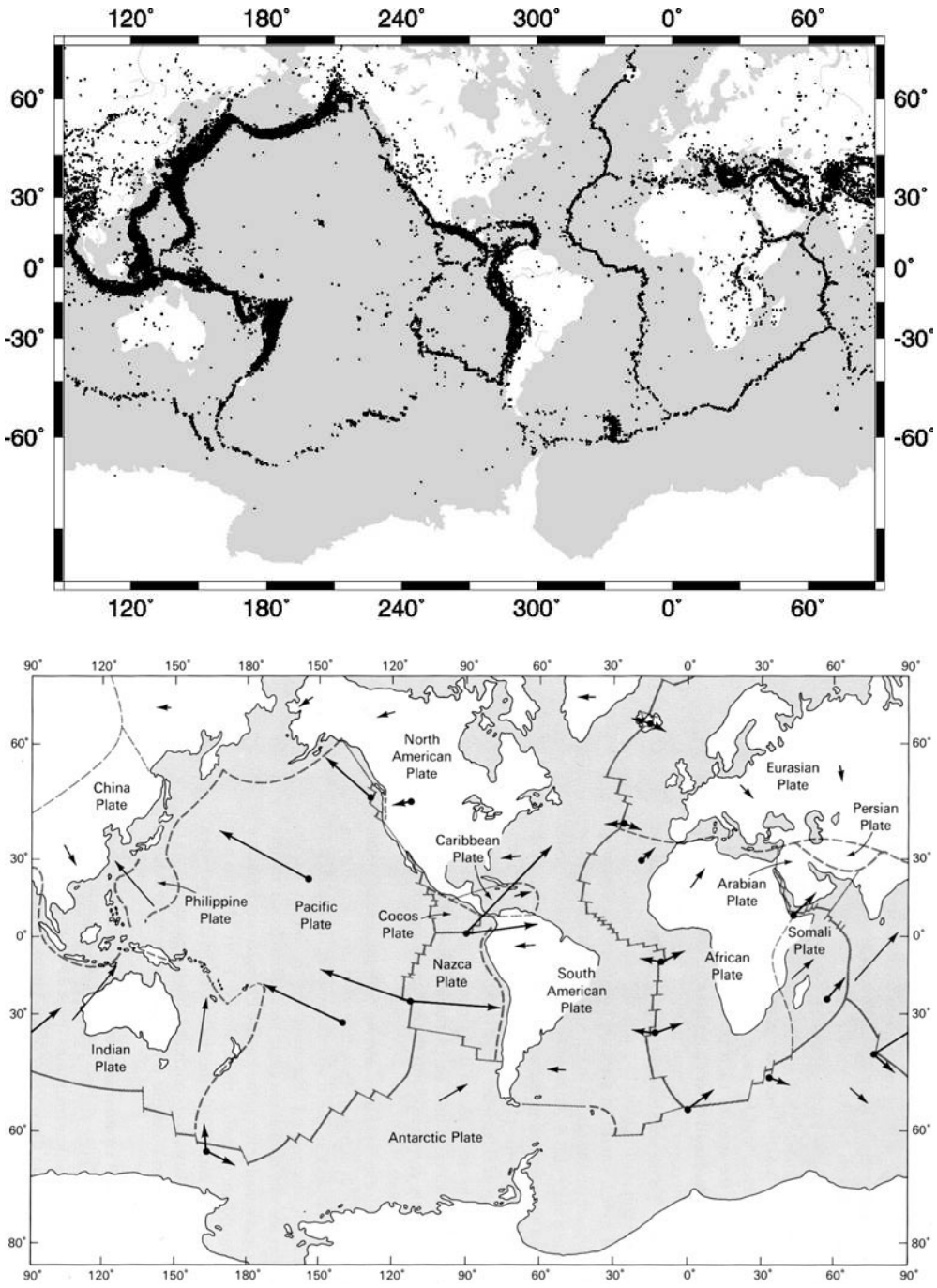

Figure 1. Global seismicity (from 1 January 1964 to 31 December 1995, magnitude range, 3.1–7.3, relocated data from the International Seismological Center catalogue) and plate motion. Earthquakes occur at the boundaries between rigid plates of the Earth's surface that move in different directions (from Uyeda (1978)).

such as satellite geodesy and high-power computation. In order to interpret the new technological advances, we must return to and push the boundaries of classical mechanical theories. The approach we take here is to emphasize the features of classical theory that are directly applicable to current, cutting-edge topics. Where possible, we highlight modern observations and laboratory results that confirm, refute or extend elements of the classical physics-based paradigm. Inevitably, our examples tend to be biased towards our own interests and research. We hope that this review will equip the reader to be properly sceptical of our results.

2. Earthquakes and stress in the crust

Earthquakes are a mechanism for accommodating large-scale motion of the Earth's plates. As the plates slide past each other, relative motion is sometimes accommodated by a relatively constant gradual slip, at rates of the order of millimetres per year; while at other times, the accumulated strain is released in earthquakes with slip rates of the order of metres per second. Sometimes, slip is accommodated by slow earthquakes or creep events with velocities of the order of centimetre per month between the two extreme cases. Current estimates are that about 80% of relative plate motion on continental boundaries is accommodated in rapid earthquakes (Bird and Kagan 2004). With few exceptions, earthquakes do not generally occur at regular intervals in time or space.

2.1. Plate motion and earthquake repeat times

The long-term loading of the Earth's crust has been traditionally measured using geodetic and geological methods. Geodesy is the branch of geophysics concerned with measuring the size and shape of the Earth's surface. The recent progress in space-based geodesy such as the Global Positioning System (GPS) and satellite interferometry (InSAR) provides us with a clear pattern of crustal movement and strain accumulation. Figure 2 shows the result of the recent geodetic measurements in Southern California. The relative plate motion determined from these data is about 2–7 cm per year which translates into a strain rate of approximately 3×10^{-7} per year along plate boundaries. The strain also accumulates in plate interiors, but at a much slower rate about 3×10^{-8} per year or less, which is an order of magnitude smaller than that at plate boundaries.

The shear strain change associated with large earthquakes (called coseismic strain drop) has been estimated using geodetic and seismological methods. For large earthquakes, it is of the order of 3×10^{-5}–3×10^{-4} (see sections 3.1.3 and 3.2.1). Since the rigidity of the crustal rocks, μ, is about 3×10^4 MPa, this corresponds to a change in shear stress (i.e. static stress drop) of about 1–10 MPa. This value is at least an order of magnitude smaller than that associated with breaking intact rocks in laboratory, which is several hundred MPa.

Dividing the coseismic strain drop by the strain rate suggests that the repeat times of major earthquakes at a given place are about 100–1000 years on plate boundaries, and 1000–10 000 years within plates. These values agree with what have been observed at many plate boundaries and interiors. This is the basic long-term process that governs global earthquake activity.

Based on the above process, a simple sketch of the stresses generating earthquakes can be drawn (figure 3(*a*)). Stress builds up on a fault plane until it reaches the breaking strength of the rock. Then, an earthquake occurs, the stress is relaxed and a new cycle begins. Although the basic process illustrated here is well understood and accurately measured, the details are more complex. For example, the loading rate is not uniform in time. A large earthquake on a segment of a fault changes the stress on the adjacent segments, either statically or dynamically,

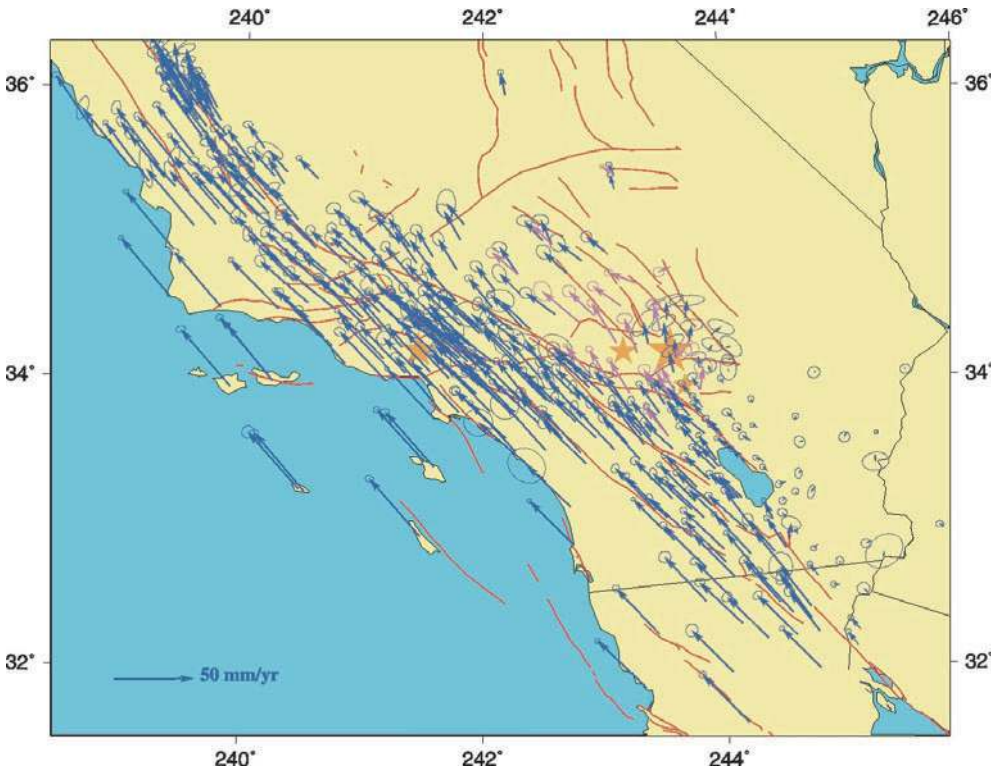

Figure 2. Velocity vectors in Southern California determined by the GPS and other space-based methods. Red lines (in the electronic version) indicate active faults. The figure is part of the Southern California Earthquake Center's web-site, http://www.scecdc.scec.org/group_e/release.v2.

and accelerates or decelerates seismic activity depending on the fault geometry. The strength of the crust is not constant in time either. Fluids may migrate in the Earth's crust, thereby weakening the crust significantly and affecting the occurrence time of earthquakes. The stress drop during earthquakes may also vary from event to event. Figure 3(*b*) illustrates these complications schematically and their effect on the intervals between earthquakes. Thus, although the overall long-term process is regular, considerable temporal fluctuations of seismicity are expected, which makes accurate prediction of earthquakes difficult.

2.2. The state of stress in the crust

As outlined earlier, the simplest model for earthquake initiation is to assume that when the stress accumulated in the plates exceeds some failure criterion on a fault plane, an earthquake happens. Evaluating this criterion requires both a measure of the resolved stress on the fault plane and a quantifiable model for the failure threshold. A first-order evaluation of the problem dates to the groundbreaking work of Anderson (1905, 1951). He started with the fact that any stress field can be completely described by its principal stresses, which are given by the eigenvectors of the stress tensor and are interpretable as the normal stresses in three orthogonal directions. He then proposed that: (1) the stress state could be resolved by assuming that one principal stress is vertical since the Earth's surface is a free surface and (2) faulting occurs when the resolved shear stress exceeds the internal friction on some plane in the medium. Internal friction is defined analogously with conventional sliding friction as a shear stress proportional to the

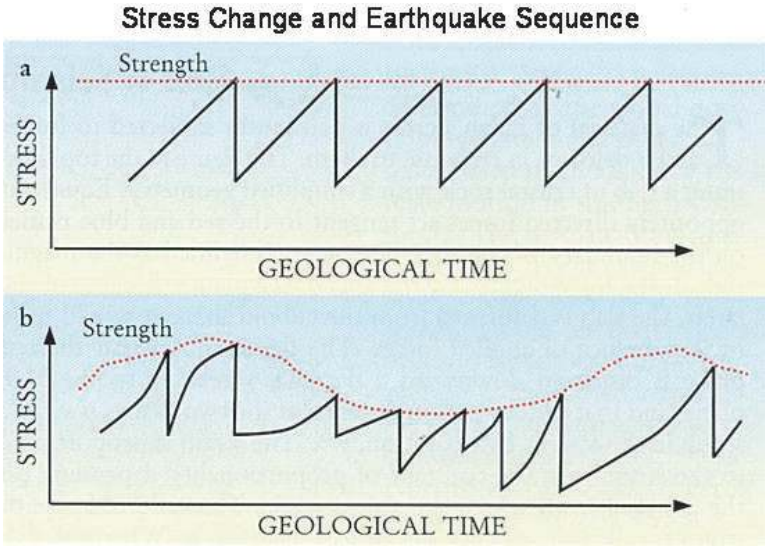

Figure 3. Stress changes and earthquake sequence. (*a*) Regular sequence. (*b*) Irregular sequence caused by the changes in loading rate and temporal variations in the strength of crust.

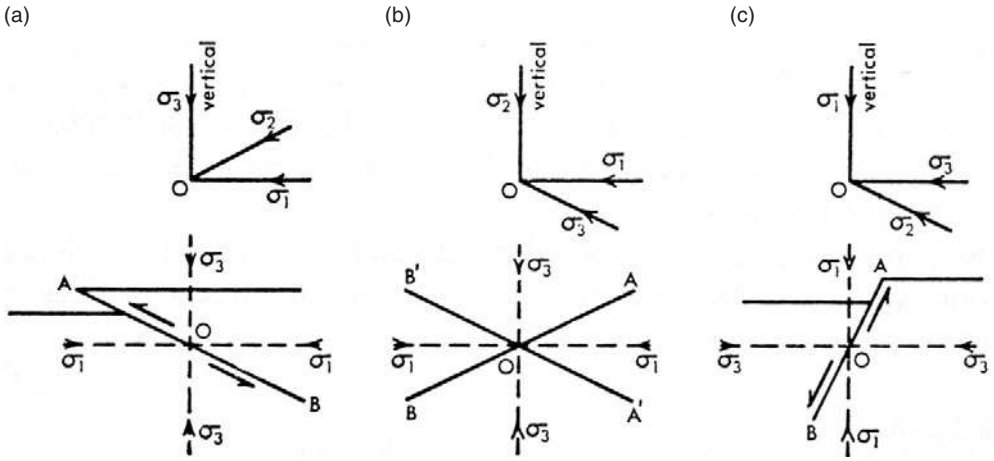

Figure 4. Schematic of the orientation of the principal stresses and the corresponding type of faulting. The principal stresses are $\sigma_1 > \sigma_2 > \sigma_3$. (*a*) Thrust faulting: the minimum principal stress is vertical. (*b*) Strike-slip faulting: the intermediate principal stress is vertical. (*c*) Normal faulting: the maximum principal stress is vertical (figure from Jaeger and Cook (1979) p 426).

normal stress on a plane. In this framework, faults are expected to accommodate horizontal motion if the vertical axis is the intermediate principal stress and accommodate both vertical and horizontal motion otherwise. A fault that has only horizontal motion is called 'strike-slip'. Combined vertical and horizontal extensional motion is called 'normal' faulting while vertical and horizontal compressional motion is called 'thrust' faulting (figure 4). Each of these three regimes corresponds to a particular orientation of the maximum principal stress.

Andersonian faulting theory has been remarkably successful in predicting and explaining the occurrence and geometry of faults. However, as we show below, a few contradictory observations cast doubt on enough parts of the paradigm that it is difficult to apply to

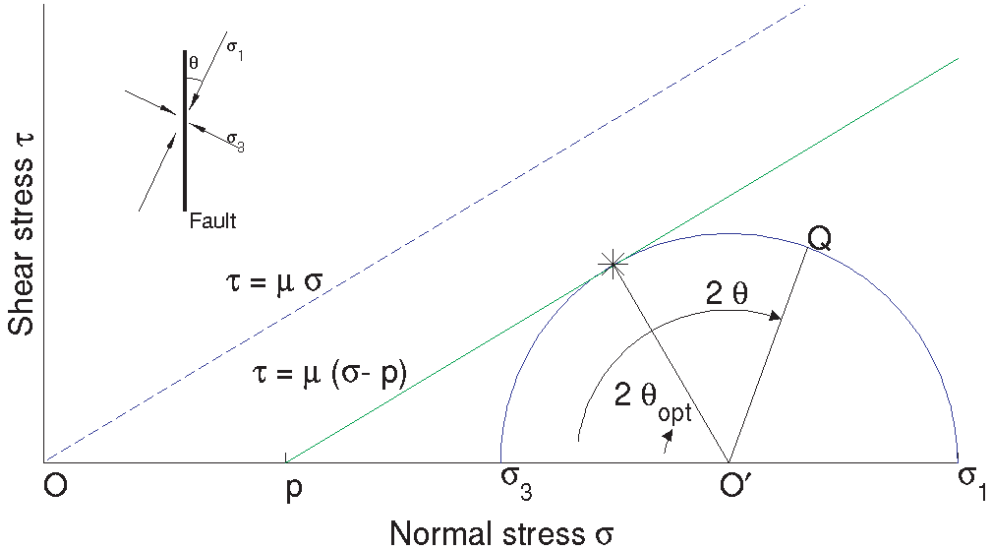

Figure 5. Mohr circle diagram. Given principal stress magnitudes σ_1 and σ_3, the locus of possible combinations of shear and normal stresses resolved on a plane are given by (2.1) and (2.2) which is plotted as the circle. The failure criterion (2.3) is the dashed line. The failure criterion in the presence of pore fluid is the solid line (2.7). Failure on a plane at an angle θ_{opt} from the orientation of σ_1 occurs if the circle intersects the failure line as it does at the $*$. Inset shows the definition of θ.

earthquakes in a straightforward way. We have difficulty measuring the coefficient of friction in the crust and have reason to believe that it varies significantly in time and space. The evidence also suggests that high fluid pressures are important in controlling frictional behaviour, yet the precise values of the ever-changing fluid pressures are also difficult to measure deep within the crust.

Principal stresses and fault orientation. Below we develop the formalism to quantitatively evaluate the frictional failure criterion in terms of the principal stresses. We will use the formalism to relate the observed geometry of faulting to the frictional strength of faults.

Denoting the principal stresses by σ_1, σ_2 and σ_3, where by definition $\sigma_1 > \sigma_2 > \sigma_3$, the relationships between the principal stresses and the resolved shear stress on a plane at an angle θ to the maximum principal stress (σ_1) can be written analytically and depicted with a Mohr circle diagram (figure 5). The convention in rock mechanics is that positive values of stresses are compressional. Since rocks are weak under tension, tensional strengths are usually $<$20 MPa, i.e. $<$10% the compressional strengths (Lockner 1995), it is generally assumed that all three principal stresses must be positive in the Earth.

The shear stress, τ, and the normal stress, σ, on the fault plane at an angle θ to σ_1 are given, respectively, by

$$\tau = \frac{\sigma_1 - \sigma_3}{2} \sin 2\theta \tag{2.1}$$

and

$$\sigma = -\frac{\sigma_1 - \sigma_3}{2} \cos 2\theta + \frac{\sigma_1 + \sigma_3}{2}. \tag{2.2}$$

A Mohr circle diagram is a plot of these two resolved stresses. The normal stress is on the x-axis and the shear stress is on the y-axis (Jaeger and Cook 1979). For a given set of principal

stresses, the solutions to equations (2.1) and (2.2) fall on a circle (figure 5). Each point on the circle represents a particular fault orientation. The angle $\angle OO'Q$ in the diagram is 2θ.

In the 17th century, Guillaume Amonton first established that the shear traction between two surfaces is proportional to the load. Amonton's Law for friction on a plane between two surfaces is written in modern terms as

$$\tau = \mu\sigma, \tag{2.3}$$

where μ is the coefficient of friction. A more complete description includes the cohesive stress C in the shear stress, i.e. $\tau = \mu\sigma + C$. However, the ratio of shear stress to normal stress, τ/σ, is more straightforward to measure, therefore most studies use an effective coefficient of friction μ which includes the cohesive effects (Lockner and Beeler 2002).

The fault planes on which slip can occur with the minimum possible deviatoric stress $\sigma_1 - \sigma_3$, i.e. the minimum diameter of the Mohr's circle, are the planes inclined at angles θ_{opt} to σ_1, such that (figure 5)

$$\tan 2\theta_{\text{opt}} = \pm\frac{1}{\mu}. \tag{2.4}$$

These two angles θ_{opt} are known as the optimal angles because they are the angles at which the rock will fracture in homogeneous, unflawed, intact rock.

Since real rocks are seldom intact, the more important criterion is the lock-up angle. If a weak plane, such as a fault, exists in the crust, the slip can be constrained to occur on that plane. In this case, for a given coefficient of friction μ on the weak plane, slip can occur at angles larger than the optimal angle θ_{opt}. However, there is a maximum value of θ beyond which slip cannot occur for any combination of positive stresses (Sibson 1985). The maximum angle, θ_{lu}, is known as the lock-up angle. From (2.1) to (2.3),

$$\frac{\sigma_1}{\sigma_3} = \frac{1 + \mu\cot\theta}{1 - \mu\tan\theta}. \tag{2.5}$$

The lock-up angle is the maximum value of θ that satisfies equation (2.5). For positive values of σ_1 and σ_3, the solution exists only if the denominator is positive, i.e. $\tan\theta \leqslant 1/\mu$. Therefore,

$$\theta_{\text{lu}} = \tan^{-1}\left(\frac{1}{\mu}\right) = 2\theta_{\text{opt}}. \tag{2.6}$$

If a fault is observed to lie at an angle θ to the maximum principal stress when it is slipping, then $\theta \leqslant \theta_{\text{lu}} = \tan^{-1}(1/\mu)$. Therefore, $\mu \leqslant \tan\theta$, and the observation gives a maximum bound on the value of μ on the fault.

Strength of the crust: laboratory and field data. Laboratory studies of rocks show that at the depths typical of earthquakes $\mu = 0.6$ to 0.85 for the majority of rocks (Byerlee 1978). Therefore, equation (2.4) predicts that faults should form at angles of 25–30° to the maximum principal stress σ_1, if they are optimally oriented. Because σ_1 is horizontal and vertical for thrust and normal faults, respectively (figure 4), the angles between the faults and the horizontal surface (i.e. dip angles) should be about 25–30° for thrust and 60–65° for normal faults if they are optimally oriented. Sibson and Xie (1998) check this criterion for the special case of intraplate thrusts. They found that 40% of the faults fall into the optimal range and none of their study sites violated the lock-up criterion. In general, only a handful of faults anywhere have been found to exceed the lock-up criterion. We will return to these unusual cases below.

The predictions of the Anderson–Byerlee mechanics have also been supported by field experiments. Boreholes are drilled and pumped full of high-pressure fluid. The pressure at which the wall of the borehole fractures and the orientation of the resulting fracture give a

measure of the magnitude and orientation of the least principal stress. More sophisticated methods use the hoop stress to infer the maximum principal stress. When these experiments are performed in an area prone to normal faulting, i.e. where the maximum principal stress is vertical, the magnitude of the stresses resultant on the fracture plane and their orientation are consistent with internal friction of 0.6 (Zoback and Healey 1984).

One complication to this simple picture was recognized early on. High fluid pressures can support part of the load across a fault and reduce the friction. In the presence of fluids equation (2.3) is modified to be

$$\tau = \mu(\sigma - p), \tag{2.7}$$

where p is the pore pressure. Hubbert and Rubey (1959) first recognized the importance of the fluid effect on fault friction. Fluid pressure at a certain depth should theoretically be determined by the weight of the water column above. This state is called hydrostatic. In the course of their work on oil exploration, Hubbert and Rubey observed that pressures in pockets of fluids in the crust commonly exceeded hydrostatic pressure. They connected this observation with studies of faulting and proposed that the pore pressure p at a depth can approach the normal stress σ on faults, resulting in low friction.

The most spectacular support for the importance of the Anderson–Byerlee paradigm of failure as modified by Hubbert and Rubey came from the 1976 Rangeley experiment. Earthquakes were induced by pumping water to increase the fluid pressure at depth in an oil field with little surface indication of faulting (Raleigh *et al* 1976). Using equations (2.1), (2.2) and (2.7), the observed fault orientation, the observed values of σ_1 and σ_3 from *in situ* borehole experiments and the measured value of μ on rock samples from the site, the researchers successfully predicted the increase in pore pressure that is necessary to trigger earthquakes.

Conflicting observations? The most controversial aspect of the Anderson–Byerlee formulation has been the applicability of the laboratory values of friction to natural settings. A fault that fails according to equation (2.7) with $\mu = 0.6$–0.85 and hydrostatic fluid pressure is called a strong fault. Three lines of evidence have complicated the Andersonian picture and led researchers to question whether or not faults are strong before and during earthquakes.

The most often cited evidence against the strong fault hypothesis is based on heat flow data. If μ is high, the frictional stress on the fault should generate heat. This heat generation, averaged over geological time should make a resolvably high level of heat flow if the depth-averaged shear stress is greater than 20 MPa. Lachenbruch and Sass (1980) showed that the San Andreas fault generates no observable perturbation to the regional heat flow pattern. Some authors have suggested that regional-scale groundwater flow may obscure such a signal, but recent modelling has shown that the data are inconsistent with any known method of removing the heat from the fault (Saffer *et al* 2003). Therefore, these difficult heat flow observations stand as the best evidence that the San Andreas has a low resolved depth-averaged shear stress ($\leqslant$20 MPa). Since this stress is lower than that which can be achieved with hydrostatic pore pressure and Byerlee friction, the fault is weak according to the definition at the beginning of this section. If the pore pressure is hydrostatic, the upper limit of 20 MPa shear stress corresponds to a maximum value of μ of 0.17. The heat flow data is sensitive only to the resolved shear stress, rather than the value of μ. Pore pressures that are more than 2.3 times the hydrostatic values can also satisfy heat flow constraint without requiring small μ. The heat flow observations can not distinguish between high pore pressure and low intrinsic fault friction.

The second line of evidence comes from geological mapping. Low-angle normal faults have now been robustly documented in the geological record (e.g. Wernicke (1981)). Although it is uncertain whether or not rapid slip occurred on these faults (as opposed to slow aseismic

creep), it is clear that large-scale movement occurred on certain faults with dip angles of 20° and perhaps as low as 2° (Axen 2004). If the faulting occurred at the lock-up angle in the more conservative case, the lock-up angle must be 70°, which translates to $\mu = 0.4$ from (2.6). Therefore, $\mu \leqslant 0.4$ on the low-angle normal faults.

Note that high pore pressure does not affect the geological result, because combining (2.6) with (2.1) and (2.2) yields

$$\frac{\sigma_1 - p}{\sigma_3 - p} = \frac{1 + \mu \cot\theta}{1 - \mu \tan\theta} \tag{2.8}$$

and the lock-up angle is still $\tan^{-1}(1/\mu)$ as long as $\sigma_3 - p \geqslant 0$. The only alternative is that p exceeds the minimum principal stress and the left-hand side (lhs) of (2.8) is negative.

A third line of evidence complicating the Anderson–Byerlee paradigm is that the maximum principal stresses next to major strike-slip faults like the San Andreas in California and Nojima in Japan are sometimes nearly normal to the fault (Zoback *et al* 1987, Ikeda 2001, Provost and Houston 2003). On the creeping zone of the San Andreas in central California, Provost and Houston find that the angle θ between σ_1 and the fault is $\sim 80°$. Therefore, according to equation (2.6) these areas must have $\mu < 0.2$ in order to be able to support motion. Further north on the fault, the angle θ varies from 40° to 70° implying a maximum value of μ varying from 0.4 to 1.2 depending on location. In Southern California, Hardebeck and Hauksson (2001) find values of θ as low as 60°. Once again, high pore pressures in the fault do not remove the need for a low value of μ in the places with high θ, if these measurements of high values of θ reflect the stress state directly on the fault. Both Byerlee (1992) and Rice (1992) argue that the stress orientation observations may not reflect the state of stress within the core of a pressurized, fluid-filled fault. If it is true that the orientations are only measured outside the fault core, then there is no constraint on the fault stress from this line of evidence.

Summary. The overall picture that is emerging is a good deal more complicated than the Andersonian view. If the framework of equations (2.1), (2.2) and (2.7) is correct then in areas with large, mature faults it appears that the μ applicable for initiation of slip must be significantly different from what is measured in the laboratory for intact rocks or immature faults like Rangely. Moreover, the stress orientation data hint that these variables may vary in time as well as space (Hardebeck and Hauksson 2001). Alternatively, pore pressure may be so high that it exceeds the minimum principal stress. However, increasing the pore pressure presents new problems as rocks can fail under tension with relatively low differential stresses. An additional complication is that μ can depend on the slip rate and its history (Dieterich 1979). Clearly, our simple criterion for earthquakes proposed above is insufficient to explain this complexity of behaviour. In order to answer our question of why earthquakes begin, we will have to dig deeper.

3. Quantifying earthquakes

In order to begin to answer these questions about earthquakes, we need to first review the major observational facts and the parameters we use to quantify earthquakes. The most developed method for measuring earthquakes is to measure the elastic wave-field generated by the sudden slip on a fault plane. Below, we discuss how the wave amplitude and frequencies are related to the physical properties of the earthquake. We then list the most common earthquake parameters derived from the wave-field and discuss their dynamical significance. Finally we explore the scaling relationships between the observed parameters.

3.1. Earthquake source parameters and observables

A formal description of the elastic problem. An earthquake is a failure process in Earth's crust. For a short-term process, we assume that the medium is elastic. We imagine that an earthquake perturbs the stress field by relaxing the stress in a localized region S embedded in the elastic medium. Prior to an earthquake, the crust is in equilibrium under some boundary conditions with the initial displacement $\vec{u}_0(\vec{r})$ and the stress distribution $\boldsymbol{\sigma}_0(\vec{r})$, where $\vec{r}$ is the position vector. The total potential energy (gravitational energy plus strain energy) of the system at this stage is W_0. In most seismological problems the displacement is assumed to be small and linear elasticity theory is used. Then, at $t = 0$, i.e. the initiation time of an earthquake, a failure occurs at a point in the medium called the earthquake hypocentre. Transient motion begins, energy is radiated, and rupture propagates into a region, S, representing the earthquake rupture zone. After the rupture propagation has stopped and the transient motion has subsided, the displacement and stress become $\vec{u}_1(\vec{r})$ and $\boldsymbol{\sigma}_1(\vec{r})$. We denote the total potential energy of this state by W_1. (Note that in section 2.2 subscripts 1, 2 and 3 are used to indicate the principal stresses; here, subscripts 0 and 1 are used to indicate the states before and after an earthquake, respectively.)

The processes in the source region S are modelled by a localized inelastic process which represents the result of the combination of brittle rupture and plastic yielding. The seismic static displacement field $\vec{u}(\vec{r})$ is

$$\vec{u}(\vec{r}) = \vec{u}_1(\vec{r}) - \vec{u}_0(\vec{r}) \tag{3.1}$$

and the stress drop is

$$\Delta\boldsymbol{\sigma}(\vec{r}) = \boldsymbol{\sigma}_0(\vec{r}) - \boldsymbol{\sigma}_1(\vec{r}). \tag{3.2}$$

The change in the potential energy is

$$\Delta W = W_0 - W_1. \tag{3.3}$$

During the failure process (i.e. coseismic process), some energy is radiated (radiated energy, E_{R}) and some energy is dissipated mechanically (fracture energy, E_{G}) and thermally (thermal energy, E_{H}). Because some parts of the fracture energy eventually become thermal energy, the distinction between E_{G} and E_{H} is model dependent.

To study an earthquake process, at least three approaches are possible.

(1) Spontaneous failure. In this case, the modelled failure growth is controlled by failure criterion (or failure physics) at each point in the medium. Thus, the final failure surface, or volume, is determined by the failure process itself. This is the most physically desirable model, but it requires the knowledge of every detail of the structure and properties of the medium. Because it is difficult to gain this information in the crust, this approach is seldom taken.

(2) Dynamic failure on a prescribed source region. In this approach, we fix the geometry of the source region. In most seismological problems, the source is a thin fault zone, and is modelled as a planar failure surface. Then what controls the rupture is the friction law on the fault plane (constitutive relation), and the elasto-dynamic equations are solved for a given friction law (often parameterized) on the fault plane. The resulting displacement field is compared with the observed field to determine the fault friction law. This approach has been taken in recent years as more computer power is available. (A recent review on this subject is given by Madariaga and Olsen (2002).)

(3) Kinematic model. In this approach, the wave-field is computed for a prescribed slip motion on the fault using the elastic dislocation theory. Then, the slip distribution on the fault is determined by the inversion of observed seismic data. At this stage, no source physics is

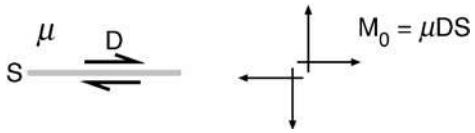

Figure 6. Representation of a dislocation (fault) seismic source. Left: a seismic fault represented by a shear displacement offset D over a surface with an area S, embedded in a medium with rigidity μ. Right: a force double couple equivalent to the dislocation model shown on left, in the limit of point source (i.e. $S \to 0$, and $D \to \infty$ while the product DS remains finite).

invoked. In this sense, this approach is called kinematic. However, once the slip is determined, it can be used to compute the associated stress field. The displacement and the stress field on the fault plane, together, can be used to infer the physical process involved in failure (i.e. friction, etc). Since many methods for inversion of seismic data have been developed, this approach is widely used. (A recent review on this subject is also given by Madariaga and Olsen (2002).)

3.1.1. Seismic source and displacement field. First, consider a very small fault (i.e. point source) on which a displacement offset D (the difference between the displacements of the two sides of a fault) occurs (figure 6, left).

We want to find a set of forces that will generate a stress field equivalent to the stress field generated by a given imposed displacement on the fault. Since the fault is entirely enclosed by elastic crust and no work is done by external forces, both linear and angular momentum must be conserved during faulting. It can be shown that the force system that respects these conservation laws and produces a stress field equivalent to the point dislocation source is the combination of two perpendicular force couples (figure 6, right). This force system is commonly called a double couple source. The moment of each force couple M_0 is given by (Stekettee 1958, Maruyama 1964, Burridge and Knopoff 1964)

$$M_0 = \mu DS, \tag{3.4}$$

where μ is the rigidity of the material surrounding the fault. (Note that in section 2.2, μ is used for the coefficient of friction, but in this section it is used to represent the rigidity. In the later sections μ is used both for the rigidity and the coefficient of friction. The distinction will be clear from the text and context.) A finite fault model can be constructed by distributing the point sources on a fault plane. The dimension of M_0 is [force] $\times$ [length] = [energy]. In seismology, it is common to use N m for the unit of M_0, rather than J (joule), because M_0 is the moment of the equivalent force system, and does not directly represent any energy-related quantity of the source.

For simplicity, we consider a homogeneous whole space with the density, ρ, the P-wave (compressional wave) velocity, α and the S-wave (shear wave) velocity, β. In the absence of any interfaces in the elastic medium, disturbances are propagated as simple elastic waves known as body waves. If a point source with seismic moment, $M_0(t)$, is placed at the origin, the displacement in the far-field is given in the polar coordinate, (r, θ, ϕ), as

$$\begin{pmatrix} u_r \\ u_\theta \\ u_\phi \end{pmatrix} = \frac{1}{4\pi\rho r\alpha^3} M_0'\left(t - \frac{r}{\alpha}\right) \begin{pmatrix} R_r(\theta,\phi) \\ 0 \\ 0 \end{pmatrix} + \frac{1}{4\pi\rho r\beta^3} M_0'\left(t - \frac{r}{\beta}\right) \begin{pmatrix} 0 \\ R_\theta(\theta,\phi) \\ R_\phi(\theta,\phi) \end{pmatrix}, \tag{3.5}$$

where the prime symbol denotes differentiation with respect to the argument.

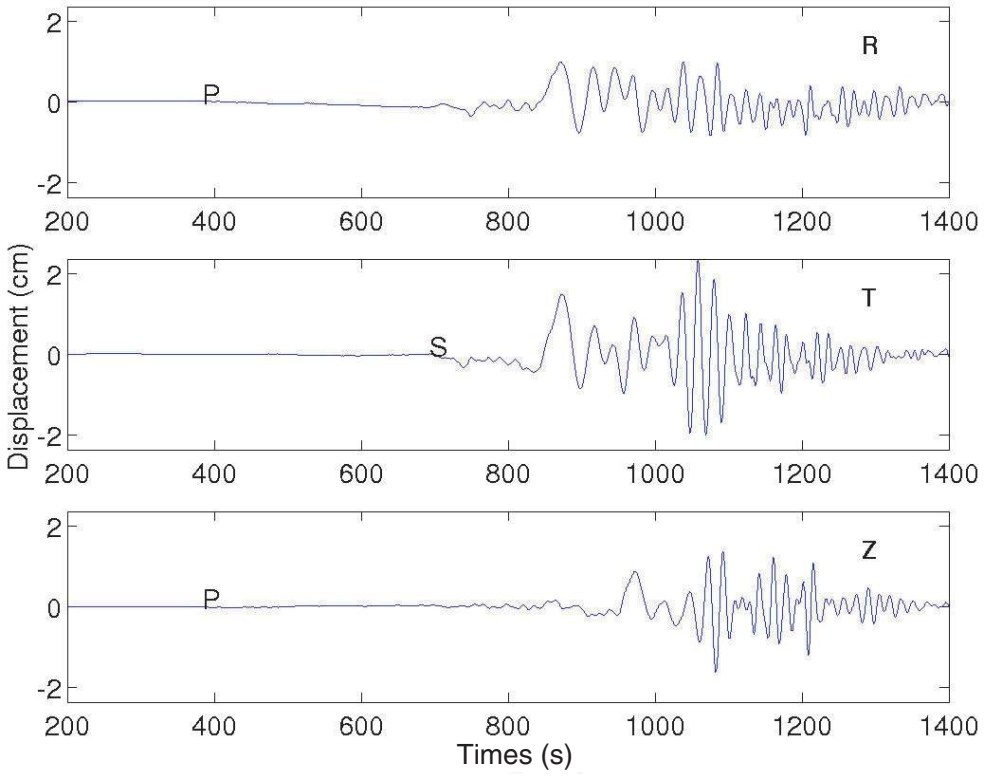

Figure 7. Example of displacement from the $M_w = 7.9$, 3 November 2002, Alaska earthquake recorded by a broadband seismometer 3460 km away in Mammoth Lakes, California. The components are radial (R), transverse (T) and vertical (Z). The radial and transverse components are the two components on the horizontal plane. The early motion on these seismograms (between 400 and 800 s) shows P- and S-waves described by (3.5). Later motion (after 800 s) shows surface waves produced by the interactions of the waves with boundaries in the earth and heterogeneous structure.

The first term is the P-wave and the second term, S-wave. $R_r(\theta, \phi)$, $R_\theta(\theta, \phi)$ and $R_\phi(\theta, \phi)$ represent the radiation patterns, which depend on the geometry of the source and the observation point. (For more details, see, e.g. Lay and Wallace (1995), Aki and Richards (2002).) These displacement components are what are measured by seismometers (figure 7).

At short distances from the source, we have an additional term representing the near-field displacement. The primary component of the near-field displacement is given approximately by

$$u \propto \frac{1}{4\pi\mu r^2} M_0(t). \tag{3.6}$$

The near-field displacement is important for the determination of detailed spatial and temporal distribution of slip in the rupture zone. Far away from the fault, (3.6) is negligible as it falls off much more quickly than the far-field terms ($1/r^2$ as opposed to $1/r$). The reason why the near-field and far-field displacements are proportional to $M_0(t)/r^2$ and $M_0'(t)/r$, respectively, is that the near-field is essentially determined by the motion on one side of the fault, while the far-field represents the contributions from both sides of the fault. This situation is similar to that of an electric field from a point charge and a dipole. (For more details, see Aki and Richards (2002)).

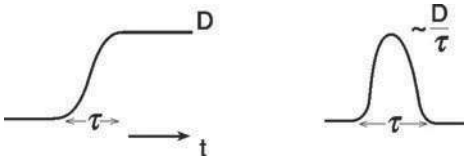

Figure 8. The near- and far-field displacements from a point dislocation seismic source which represents a fault slip motion given by a ramp function with duration τ.

If the fault motion is a linear ramp function, then $M_0(t)$ is a ramp function, which, after differentiated, produces a box-car far-field wave form. In general, if the fault motion occurs over a duration of τ, then the near-field wave form is a ramp function and the far-field wave form is a pulse with a duration of τ (figure 8).

The time derivative of the seismic moment $\dot{M}_0(t)$ is called the moment-rate function or the source time function, and its frequency spectrum is called the moment rate spectrum or the source spectrum.

So far, we have discussed the seismic body waves that travel directly from the earthquake to the seismometer. In the real situation, seismic body waves interact with the Earth's free surface and many internal structural boundaries to develop reflections as well as surface waves (i.e. the wave trains in the later parts of figure 7). Observed surface waves are often very long period, 10–300 s, and carry the information of the seismic source at long period. When surface waves propagate around the Earth many times, they can be interpreted as elastic oscillations of the Earth. The long-period oscillations can be studied using the normal mode theory. The theories of seismic surface waves and normal modes are well developed, and have been used effectively to study earthquakes (Gilbert and Dziewonski 1975, Dahlen and Tromp 1998).

3.1.2. Seismic moment and magnitude. We now use the elastic theory developed above to determine parameters of earthquakes that measure the size, energy and stress during rupture.

As shown by (3.5) and (3.6), the seismic moment can be determined from the integral of the far-field displacement, or from the amplitude of the near-field displacement. In the actual determination of the seismic moment, we need to include the effect of wave propagation in a heterogeneous structure, geometry of the source and the finiteness of the source. Many seismological methods have been developed to handle these problems, and the seismic moment can be determined accurately from seismic data (e.g. Lay and Wallace (1995)). For a finite source with a fault area S on which the spatially averaged slip is $\bar{D}$ (offset), the seismic moment M_0 is given by $\mu \bar{D} S$. Because M_0 depends on the two end states, before and after an earthquake, it does not depend on the actual time history of faulting. In this sense, it is a static parameter. If M_0 is determined by a seismic method, and if S is estimated by either a seismic or geodetic method, $\bar{D}$ can be determined by using the relation $\bar{D} = M_0/\mu S$.

The seismic moment M_0 can be determined:

(1) From seismic data: the amplitude of long-period surface waves and normal-modes can be used to determine M_0 most accurately, because long-period waves are least affected by complex propagation effects. The amplitude and frequency spectra of seismic body waves can be also used for smaller earthquakes. This is the most common method.

(2) From geodetic data: with the advent of space-based geodetic methods (e.g. GPS and InSAR), this method is becoming more commonly used. The synthetic aperture radar (SAR) interferometry was used for the 1992 Landers, California, earthquake (Massonnet *et al* 1993)

Table 1. Seismic moment determinations from different data sets.

Data	M_0 (N m)	Reference
Hector Mine, California, Earthquake, 16 October 1999, $M_w = 7.1$		
Long-period surface waves	5.98×10^{19}	Harvard University
Seismic body waves	5.5×10^{19}	Earthquake Research Institute, Tokyo University
GPS and InSAR	6.7×10^{19}	Simons *et al* (2002)

to successfully map the regional static deformation field associated with this earthquake. To determine M_0 accurately, good spatial coverage around the source is required.

(3) From geological data: the surface break of a fault can be used to estimate M_0. However, the distribution of slip where a fault meets the surface of the Earth does not necessarily represent the slip at depths, and the resulting estimate of M_0 is inevitably inaccurate. However, for historical events for which no instrumental data are available, this method is often used.

The redundant multiple methods allow us to verify that seismic moment is well-measured by seismic methods to an accuracy unequalled by any other seismic parameters. Table 1 shows the results for the 1999 Hector Mine, California, earthquake where the seismic moment was independently measured by methods 1 and 2. The values determined by different methods generally agree within 30%.

The following web-sites provide a catalogue of seismic moment of large earthquakes in the world, compiled by the Seismology Group of Harvard University, Earthquake Information Center of the Earthquake Research Institute of Tokyo University and the United States Geological Survey, respectively.

- http://www.seismology.harvard.edu
- http://wwweic.eri.u-tokyo.ac.jp/EIC/EIC_News/index-e.html
- http://neic.usgs.gov/neis/FM/previous_mom.html

Seismic moments are the most modern and accurate quantification of the size of an earthquake; however, historically, magnitude scales were used for this purpose. Most magnitude scales were defined by the observed amplitude of seismic waves with some corrections for attenuation with distance from the source, but these magnitudes are empirical parameters and cannot be directly related to any specific physical parameter of the source. Recently, the standard practice is to define the magnitude with the seismic moment. This magnitude, M_w, is defined by the following relation:

$$M_w = \frac{\log_{10} M_0}{1.5} - 6.07 \qquad (M_0 \text{ in N m}). \tag{3.7}$$

As mentioned above, M_0 is a static parameter and does not represent any dynamic properties of the source. However, with the use of some scaling relations, it can be approximately related to the total radiated energy, at least for large earthquakes (section 3.2.2). In this sense, M_0 or M_w can be used as a useful quantification parameter for an earthquake and its damaging effects.

3.1.3. Strain and stress drop. As we discussed above, the stress drop caused by an earthquake is $\Delta\boldsymbol{\sigma}(\vec{r}) = \boldsymbol{\sigma}_0(\vec{r}) - \boldsymbol{\sigma}_1(\vec{r})$. We usually consider only the shear stress on the fault plane,

$$\Delta\sigma_s = \sigma_0 - \sigma_1 \tag{3.8}$$

and call it the static stress drop associated with an earthquake. The strain drop $\Delta\varepsilon_s$ is given by $\Delta\varepsilon_s = \Delta\sigma_s/\mu$. In general, $\Delta\sigma_s$ varies spatially on the fault. The spatial average is given by

$$\overline{\Delta\sigma_s} = \frac{1}{S}\int_S \Delta\sigma_s \, dS. \tag{3.9}$$

Since the stress and strength distributions near a fault are non-uniform, the slip and stress drop are, in general, complex functions of space. In most applications, we use the stress drop averaged over the entire fault plane. The stress drop can be locally much higher than the average. To be exact, the average stress drop is the spatial average of the stress drop, as given by (3.9). However, the limited resolution of seismological methods often allows determinations of only the average displacement over the fault plane, which in turn is used to compute the average stress drop. With this approximation, we estimate $\overline{\Delta\sigma_s}$ simply by

$$\overline{\Delta\sigma_s} \approx C\mu\frac{\bar{D}}{\tilde{L}}, \tag{3.10}$$

where, $\bar{D}$ is the average slip (offset), $\tilde{L}$ is a characteristic rupture dimension, often defined by $\sqrt{S}$ and C is a geometric constant of order unity. Unfortunately, given the limited spatial resolution of seismic data, we cannot fully assess the validity of this approximation. However, Madariaga (1977, 1979), Rudnicki and Kanamori (1981) and Das (1988) show that this is a good approximation unless the variation of stress on the fault is extremely large.

We often use $\Delta\sigma_s$ to mean the average static stress drop in this sense. Some early determinations of stress (strain) drops were made using D and $\tilde{L}$ estimated from geodetic data (e.g. 1927 Tango earthquake, Tsuboi (1933)).

More commonly, if the seismic moment is determined by either geodetic or seismological methods, we use the following expression. Using $M_0 = \mu\bar{D}S$, $\tilde{L} = S^{1/2}$ and (3.10), we can write

$$\overline{\Delta\sigma_s} = CM_0\tilde{L}^{-3} = CM_0S^{-3/2}. \tag{3.11}$$

If the length scale of the source is estimated from the geodetic data, aftershock area, tsunami source area or other data, we can estimate the stress drop using (3.11) (e.g. Kanamori and Anderson (1975), Abercrombie and Leary (1993)).

If the slip distribution on the fault plane can be determined from high-resolution seismic data, it is possible to estimate the stress drop on the fault plane (Bouchon 1997).

Since $\overline{\Delta\sigma_s} \approx CM_0\tilde{L}^{-3}$, an uncertainty in the length scale can cause a large uncertainty in $\overline{\Delta\sigma_s}$: a factor of 2 uncertainty in $\tilde{L}$ results in a factor of 8 uncertainty in $\overline{\Delta\sigma_s}$. Thus, an accurate determination of earthquake source size, either S or $\tilde{L}$, is extremely important in determining the stress drop.

3.1.4. Energy

Radiated energy, E_R. The energy radiated by seismic waves, E_R, is another important physical parameter of an earthquake. In principle, if we can determine the wave-field completely, it is straightforward to estimate the radiated energy. For example, if the P-wave displacement in a homogeneous medium is given by $u_r(r, t)$, then the energy radiated in a P-wave is given by

$$E_{R,\alpha} = \rho\alpha\int_{S_0}\int_{-\infty}^{+\infty} \dot{u}_r(r,t)^2 \, dt \, dS_0, \tag{3.12}$$

where S_0 is a spherical surface at a large distance surrounding the source. Similarly, the energy radiated in an S-wave is given by

$$E_{R,\beta} = \rho\beta\int_{S_0}\int_{-\infty}^{+\infty} [\dot{u}_\theta(r,t)^2 + \dot{u}_\phi(r,t)^2] \, dt \, dS_0. \tag{3.13}$$

Table 2. Determinations of radiated energy with different data sets and methods.

Data	E_R (J)	Reference
Bhuji, India, Earthquake, 26 January 2001, $M_w = 7.6$		
Regional data	2.1×10^{16}	Singh *et al* (2004)
Teleseismic data	2.0×10^{16}	Venkataraman and Kanamori (2004)
Frequency-domain method	1.9×10^{16}	Singh *et al* (2004)
Hector Mine, California, Earthquake, 16 October 1999, $M_w = 7.1$		
Regional data	3.4×10^{15}	Boatwright *et al* (2002)
	3×10^{15}	Venkataraman *et al* (2002)
Teleseismic data	3.2×10^{15}	Boatwright *et al* (2002)
	2×10^{15}	Venkataraman *et al* (2002)

The total energy, E_R, is the sum of $E_{R,\alpha}$ and $E_{R,\beta}$ (e.g. Haskell (1964)). In practice, however, the wave-field in the Earth is extremely complex because of the complexity of the seismic source, propagation effects, attenuation and scattering. Extensive efforts have been made in recent years to accurately determine E_R. For earthquakes for which high-quality seismic data are available, E_R can be estimated probably within a factor of 2–3 (McGarr and Fletcher 2002). Some examples are shown in table 2.

Potential energy. The potential energy change in the crust due to an earthquake is

$$\Delta W = \tfrac{1}{2}\overline{(\sigma_0 + \sigma_1)D}S, \tag{3.14}$$

where the bar stands for the spatial average (Kostrov 1974, Dahlen 1977). Equation (3.14) can be rewritten as,

$$\Delta W = \tfrac{1}{2}\overline{(\sigma_0 - \sigma_1)D}S + \overline{\sigma_1 D}S = \tfrac{1}{2}\overline{\Delta\sigma_s D}S + \overline{\sigma_1 D}S = \Delta W_0 + \overline{\sigma_1 D}S, \tag{3.15}$$

where

$$\Delta W_0 = \tfrac{1}{2}\overline{\Delta\sigma_s D}S. \tag{3.16}$$

Two difficulties are encountered. First, with seismological measurements alone, the absolute value of the stresses, σ_0 and σ_1 cannot be determined. Only the difference $\Delta\sigma_s = \sigma_0 - \sigma_1$ is determined. Thus, we cannot compute ΔW from seismic data. As we discussed in section 2.2, non-seismological methods give inconsistent results for background stress. Second, as we discussed in section 3.1.3 for the stress drop, with the limited resolution of seismological methods, the details of spatial variation of stress and displacement cannot be determined. Thus we commonly use, instead of (3.16),

$$\Delta W_0 = \tfrac{1}{2}\overline{\Delta\sigma_s}\ \overline{D}S. \tag{3.17}$$

Unfortunately, it is not possible to accurately assess the errors associated with the approximation of equation (3.17). It is a common practice to assume that the approximation is sufficiently accurate if the spatial variation is not very rapid.

Although ΔW cannot be determined by seismological methods, ΔW_0 can be computed from the seismologically determined parameters, $\overline{\Delta\sigma_s}$, $\overline{D}$ and S. In general $\overline{\sigma_1 D}S > 0$, unless a large scale overshoot occurs, and ΔW_0 can be used as a lower bound of ΔW. If the residual stress σ_1 is small, ΔW_0 is a good approximation ofΔW.

It is important to note that we can determine two kinds of energies, the radiated energy (E_R) and the lower bound of the potential energy change (ΔW_0), with seismological data and methods. These two energies play an important role in understanding the physics of earthquakes (section 4.4).

3.1.5. Rupture mode, speed and directivity. Another observable feature of earthquakes is the rupture pattern on the fault. Although the rupture pattern is not a parameter *sensu stricto*, since it is not a single summary quantity, it is another observable characterization of the rupture. From the rupture patterns we can define some secondary parameters describing rupture propagation velocity, slip duration and directivity.

An earthquake occurs on a finite fault. It initiates from a point, called a hypocentre, and propagates outward on the fault plane. From the gross rupture patterns, we classify the rupture patterns into unilateral rupture, bilateral rupture and two-dimensional (approximately circular) rupture patterns. In a unilateral rupture, the hypocentre is at the one end of the fault, and the rupture propagates primarily in one direction toward the other end. One good example is the recent Denali, Alaska, earthquake ($M_w = 7.9$, 3 November 2002). In a bilateral rupture, the rupture propagates in opposite directions from the hypocentre. Good examples include the 1989 Loma Prieta, California, earthquake ($M_w = 6.9$), and the 1995 Kobe, Japan, earthquake ($M_w = 6.9$). Bilateral ruptures are not necessarily symmetric. The 1906 San Francisco, California, earthquake is believed to have ruptured in both directions, but propagated further to the north than the south. In the description of unilateral and the bilateral fault, the fault geometry is assumed to be one-dimensional. In some earthquakes, the rupture propagates in all directions on the fault plane. In these cases, a circular fault is often used to model the fault plane.

Directivity and source duration. The source finiteness and rupture propagation have an important effect on seismic radiation. This effect, called directivity, is similar to the Doppler effect.

As we discussed in section 3.1.1, the far-field displacement is given by the time derivative of the near-field displacement and is pulse-like. However, because of the Doppler effect, the observer located in the direction of the rupture will see a shorter pulse than the observer in the direction away from the rupture direction. However, the area under the pulse-like waveform (i.e. the displacement integrated over time) is constant, regardless of the azimuth and is proportional to the seismic moment, M_0. The duration of the pulse, $\bar{\tau}$, when averaged over the azimuth, is proportional to the length scale of the fault $\tilde{L}$ divided by the rupture speed, V

$$\bar{\tau} = \frac{\tilde{L}}{V}. \tag{3.18}$$

For unilateral, bilateral and circular faults, $\tilde{L}$ is commonly taken to represent the fault length, half the fault length and the radius of the fault plane, in that order. The variation of the pulse width as a function of azimuth due to rupture propagation has an important influence on ground motions. As mentioned above, at a site towards the rupture propagation, the pulse becomes larger and narrower, and produces stronger ground motions which often result in heavier shaking damage, than at a site away from the rupture propagation. A good example is the 1995 Kobe, Japan, earthquake. The rupture of one of the bilateral segments propagated northeast from the hypocentre toward Kobe and produced very strong ground motions in the city.

Rupture speed. As we will discuss later (section 4.6.1), the rupture speed is an important parameter which reflects the dynamic characteristic of a fracture. In particular, the fraction of the shear velocity that the shear crack rupture velocity achieves is related to the fracture energy. Thus it is important to determine the rupture speed of earthquake faulting to understand the nature of earthquake mechanics.

The rupture speed V has been determined for many large earthquakes. In general, for most large shallow earthquakes, V is approximately 75–95% of the S-wave velocity at the depth where the largest slip occurred. However, there are some exceptions. For some earthquakes

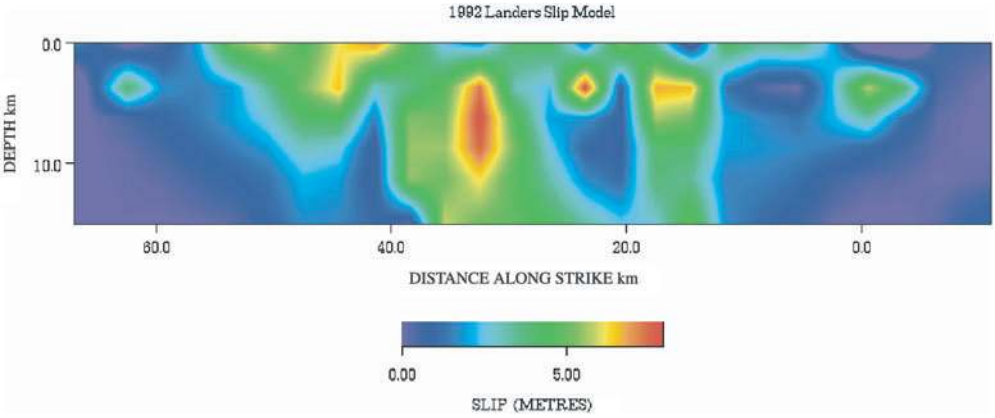

Figure 9. Complexity of earthquake rupture pattern. Rupture pattern for the 1992 Landers, California, earthquake determined with inversion of seismic data (Wald and Heaton 1994).

(e.g. 2001 Kunlun, China, earthquake), super-shear rupture velocities, i.e. $V > \beta$, have been reported (Bouchon and Vallee 2003). For some earthquakes (e.g. 1992 Nicaragua earthquake, Kikuchi and Kanamori (1995)), a very slow rupture speed has been reported. For deep earthquakes, an accurate determination of V is usually difficult, because of the difficulty in resolving the rupture pattern due to the lack of close-in observations. For the largest deep earthquake, the 1994 Bolivian earthquake ($M_w = 8.3$), the resolution of the seismic method was good enough to determine V; a very low, $(V/\beta) = 0.2$ (e.g. Kikuchi and Kanamori (1994)), rupture speed has been reported. For other smaller deep earthquakes, higher rupture speeds have been reported (e.g. Tibi *et al* (2003)).

The relatively high rupture speeds observed for most shallow earthquakes is in striking contrast with the rupture speeds observed in laboratory. Most of the laboratory data show that the rupture speed for intact materials under tensile stress is at most 50% of the Rayleigh wave speed. It is not possible to maintain a shear fault in intact materials, because the rupture bifurcates and cannot produce a planar faulting. Higher rupture speeds have only been observed for pre-cut samples. In a few pre-cut experiments the rupture velocities are even higher than shear wave speed (Rosakis 2002).

The difference and similarity between earthquakes and laboratory fractures provide an important clue to the mechanics of earthquake rupture, as we will discuss in section 4.6.1.

3.1.6. Earthquake rupture pattern. The slip distribution in real earthquakes is very complex. With the advent of modern strong-motion seismographs and broad-band seismographs, it has become possible to determine the actual slip distribution by inverting the observed seismic waveforms. These studies demonstrate that the slip distribution on a fault plane is highly heterogeneous in space and time, as shown for the 1992 Landers, California, earthquake (figure 9, Wald and Heaton (1994)). However, in most modelling studies, short-period (usually 2 s or shorter) waves are filtered out because of the difficulty in modelling such short-period waves. At periods shorter than 2 s (the corresponding wavelength is about $\lambda = 5$ km), scattering of waves and complexities of the source process produce wave forms too complex to be explained with a simple model. Thus, these models should be regarded as long-wavelength rupture patterns; the real slip distribution is probably far more complex with short wavelength irregularities. Although the spatial resolution of these models are not always given, it is probably of the order of $\lambda/3$. Short wavelength heterogeneity has been suggested by

complex high-frequency wave forms seen on accelerograms recorded at short distances. Zeng *et al* (1994) modelled an earthquake fault as a fractal distribution of patches. This complexity suggests that the microscopic processes on a fault plane can be important in controlling the rupture dynamics, as we will discuss in section 4.5.

3.2. Seismic scaling relations

Now that we have measured some average properties of rupture, we need to relate the parameters to each other. The scaling relationships between the macroscopic source parameters are useful for isolating general constraints on the microscopic forces and processes in the fault zone during rupture. We will first discuss a selection of the observed scalings with only a cursory overview of the implications. A more detailed discussion of microscopic physics follows in section 4.

3.2.1. Scaling relations for static parameters. The seismic moment, M_0, is the source parameter that can be determined most reliably. Thus, it is useful to investigate a scaling relation between M_0 and another parameter that can be determined most directly from seismic observations.

(1) M_0 versus source duration. We first choose the duration of source process, τ. This parameter can be determined from a seismogram, but it is not just the duration of a waveform recorded on a seismogram. We must remove the propagation effects from the seismograms to estimate the duration of rupture process at the source. τ is equal to the azimuthal average of $\bar{\tau}$ discussed in section 3.1.5, equation (3.18). The existing data (Masayuki Kikuchi, written communication (2001)) show a gross scaling relation

$$M_0 \propto \tau^3, \tag{3.19}$$

as shown in figure 10.

(2) Moment versus fault area. It is not always easy to determine the fault area S (i.e. rupture area), but by combining various kinds of data (e.g. aftershock area, surface rupture, geodetic data, directivity and seismic inversion results), the rupture areas for large ($M_w > 6$) earthquakes have been estimated. Figure 11 shows the results, and suggests a scaling relation

$$M_0 \propto S^{3/2}. \tag{3.20}$$

The scaling relation given by (3.20) can be interpreted as follows. From equation (3.11), the seismic moment M_0, $\overline{\Delta\sigma_s}$, and the length scale $S^{1/2}$ are related by

$$M_0 = \frac{1}{C}\overline{\Delta\sigma_s} S^{3/2}. \tag{3.21}$$

Hereafter $\overline{\Delta\sigma_s}$ is simply written as $\Delta\sigma_s$ for brevity. If $\Delta\sigma_s$ is constant, then $M_0 \propto S^{3/2}$, which is the scaling relation shown in figure 11. Thus, this scaling relation indicates that $\Delta\sigma_s$ is roughly constant over a range of M_0 from 10^{18} to 10^{23} N m. The level of the curve determines the value of $\Delta\sigma_s$. From figure 11 we can estimate that $\Delta\sigma_s$ is, on the average, approximately 3 MPa with a range 1–10 MPa. Because of the uncertainty in S, and the assumption for the geometry of the fault plane, this estimate of $\Delta\sigma_s$ is not precise, but the approximate range 1–10 MPa is considered robust. There is some evidence that $\Delta\sigma_s$ varies for different types of earthquakes, such as those on major plate boundaries and those in plate interiors, but the overall difference is probably within this range. Figure 11 is the evidence that most earthquakes have comparable stress drops in the range 1–10 MPa .

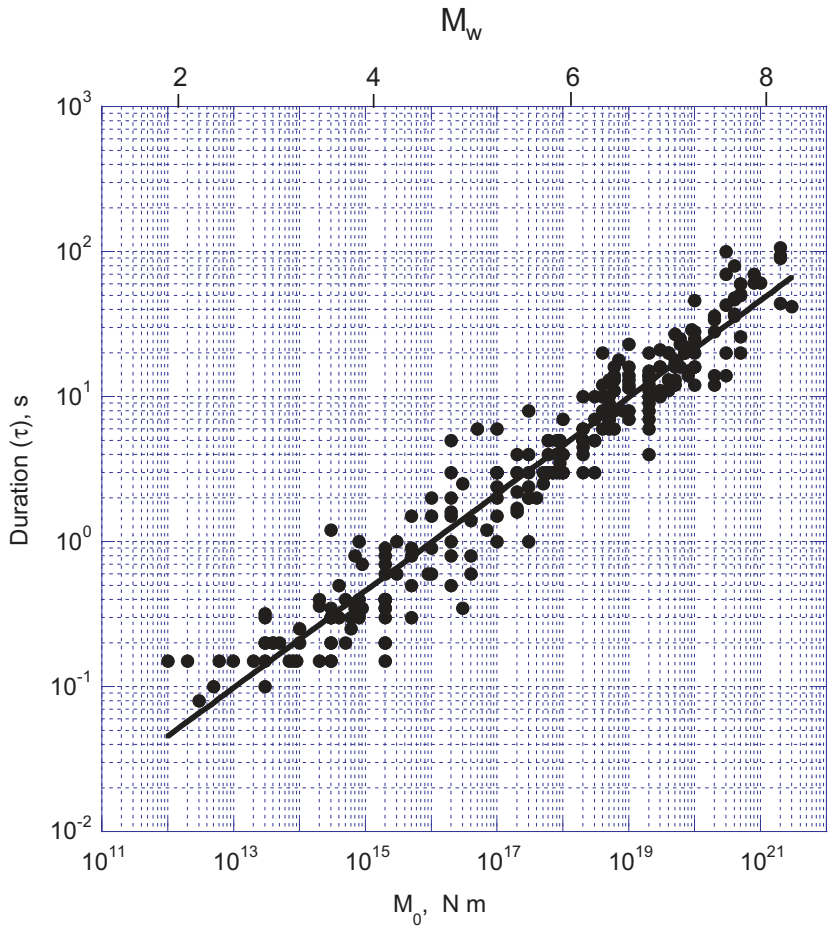

Figure 10. The relation between the seismic moment M_0 (the corresponding M_w is shown at the top of the figure) and the source duration τ for shallow (depth <60 km) earthquakes. The data are for events during the period from 1991 to 2001. The data for events larger than $M_w = 6.5$ ($M_0 = 7 \times 10^{18}$ N m) are for the events worldwide. The data for smaller events in Japan are added. (Masayuki Kikuchi, written communication (2001)). The horizontal and vertical alignments of the data points are an artefact due to the rounding-off of the values used for M_0 and τ.

Comparison of this scaling relation, $M_0 \propto S^{3/2}$, with the scaling relation, $M_0 \propto \tau^3$ (3.19) suggests that $\tau \propto S^{1/2}$. We define the length scale of the fault to be $\tilde{L} \equiv S^{1/2}$. Since $\tilde{L} \approx V\tau$ (3.18) where V is the rupture speed, this means that V is constant for most shallow earthquakes. As mentioned earlier, for some shallow large earthquakes, the rupture speed V is directly determined to be 75–95% of the S-velocity. Thus, these results, taken together, suggest that most large shallow earthquakes have $\Delta\sigma_s$ ranging from 1 to 10 MPa, and the rupture speed V is roughly constant at 75–95% of the S-velocity. We note here that these are the general scaling relations, and there are exceptions.

A similar analysis can be made for smaller earthquakes. However, it is difficult to determine the source dimension of small (e.g. $M_w < 3$) earthquakes directly. In most cases, the pulse width or the corner-frequency of the source spectrum is used to infer the source dimension. (As discussed in section 3.1.5, the pulse width is, on the average, equal to L/V.

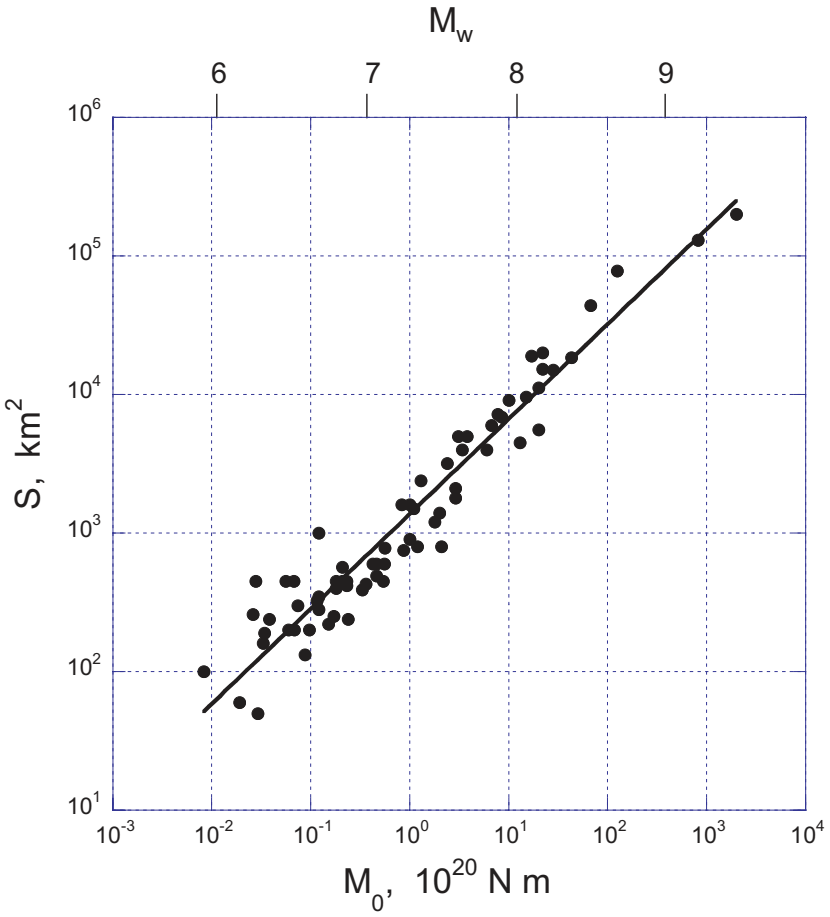

Figure 11. The scaling relation between seismic moment M_0 (the corresponding M_w is shown at the top of the figure) and the fault area S for shallow earthquakes. The data are from Kanamori and Anderson (1975) and Masayuki Kikuchi (written communication 2001). The unpublished data provided by Kikuchi are for the events worldwide during the period 1991 to 2001 for which the source dimension could be estimated.

If the rupture speed, V, is approximately equal to the S-velocity, the pulse width can be used to estimate the fault length. The corner frequency of the spectrum of a pulse-like source function is proportional to the reciprocal of the pulse width.) The general trend follows the $M_0 \propto S^{3/2}$ scaling, with $\Delta\sigma_s$ ranging from 0.1 to 100 MPa (Abercrombie 1995). This large range in $\Delta\sigma_s$ may be real, reflecting the heterogeneities of the crust on short length scales. It is also possible that the large scatter is due to errors in determining the source dimension. At present, this question is not resolved.

In the scaling relations discussed earlier, the length scale of the source is defined by $S^{1/2}$ with the idea of representing the source dimension with just one parameter. However, different faults have different aspect ratios (i.e. the ratio of fault length to width). For example, for long crustal strike-slip earthquakes such as the 1906 San Francisco earthquake, the fault length is about 350 km, but the depth-wise width of the fault is probably comparable to the upper half of the crust, about 15 km. In contrast, large subduction-zone earthquakes like the 1964

Alaskan earthquake have a fault width as large as 200 km or more. In view of this variation in aspect ratio, several investigators tried to investigate the scaling relation between M_0 and L (e.g. Romanowicz and Rundle (1993), Scholz (1994), Romanowicz and Ruff (2002)). Several different scaling relations, such as $M_0 \propto L^{3/2}$ and $M_0 \propto L^2$, have been proposed for different types of earthquakes and for different magnitude ranges.

3.2.2. Scaling relations for dynamic parameters. The radiated energy, E_R, is another macroscopic earthquake source parameter that can be determined by seismological methods (see section 3.1.4). The ratio

$$\tilde{e} = \frac{E_R}{M_0} = \frac{1}{\mu}\frac{E_R}{\bar{D}S} \tag{3.22}$$

has long been used in seismology as a useful parameter that characterizes the dynamic properties of an earthquake (Aki 1966, Wyss and Brune 1968). The ratio $\tilde{e}$ multiplied by the rigidity μ is called the apparent stress. From (3.22), the ratio can be interpreted as proportional to the energy radiated per unit fault area and per unit slip. In this sense, this scaling relation represents a dynamic property of earthquakes. As we will discuss later (section 4.4), if the static stress drop is constant, then $\tilde{e}$ must be constant if small and large earthquakes are dynamically similar. Seismologists are very concerned with whether or not earthquakes are dynamically similar because of the implications of the observation for the predictability of the eventual size of an earthquake. If small and large earthquakes are dynamically similar, then the initiation process is scale-invariant and therefore the size of earthquakes is inherently unpredictable. However, the converse statement is not true, so the observation of a lack of similarity cannot prove the predictability of earthquake size.

In view of its importance for understanding the dynamic character of earthquakes, many studies have been devoted to the determination of $\tilde{e}$. Unfortunately, it is difficult to determine E_R accurately, because of the complex wave propagation effects in the Earth, especially for small earthquakes, and the results were widely scattered.

Recent improvements in data quality and methodology have significantly improved the accuracy of E_R determination for large earthquakes (e.g. $M_w > 6$) (e.g. Boatwright and Choy (1986), Boatwright *et al* (2002), Venkataraman *et al* (2002)). For small earthquakes, it is still difficult because the relatively high-frequency seismic waves excited by small earthquakes are easily scattered and attenuated by the complex rock structures between the fault and a seismic station. Nevertheless, using down-hole instruments, or with the careful removal of path effects, large amounts of high-quality data for small earthquakes have been accumulated (Abercrombie 1995, Mayeda and Walter 1996, Izutani and Kanamori 2001, Prejean and Ellsworth 2001, Kinoshita and Ohike 2002). The ratio depends on many seismogenic properties of the source region so that it varies significantly for earthquakes in different tectonic environments, such as continental crust, subduction zone, deep seismic zone, etc (Choy and Boatwright 1995, Perez-Campos and Beroza 2001, Venkataraman and Kanamori 2004). However, the data for the same type of earthquakes exhibit an interesting trend. Figure 12 shows the results for crustal earthquakes in California and Japan.

Taken at face value, despite the large scatter, the average ratio $\tilde{e}$ decreases as the magnitude, M_w, decreases. For large earthquakes ($M_w \approx 7$), $\tilde{e}$ is, on the average, approximately 5×10^{-5}, but it is approximately a factor of 10 smaller at $M_w \approx 3$, and a factor of 100 smaller at $M_w \approx 1$. Results for even smaller earthquakes show even smaller values of $\tilde{e}$ (e.g. Jost *et al* (1998), Richardson and Jordan (2002)). Ide and Beroza (2001) suggested that many of the published $\tilde{e}$ versus M_w relations could be biased to have decreased $\tilde{e}$ for small events because of inadequate corrections for path effects or the limited instrumental pass-band. These systematic

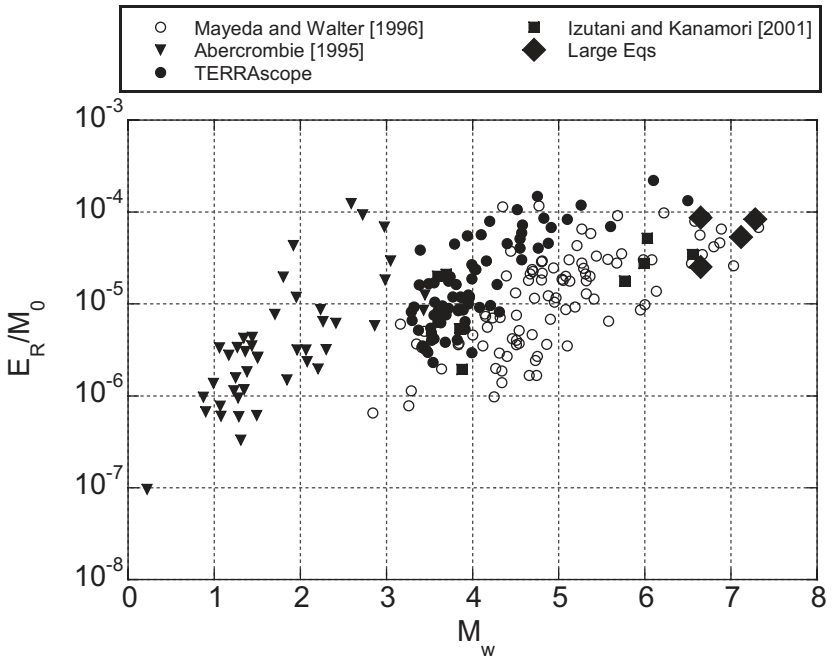

Figure 12. The relation between $\tilde{e} = E_R/M_0$ and M_w (Abercrombie 1995, Mayeda and Walter 1996, Izutani and Kanamori 2001, Kanamori *et al* (1993), for TERRAscope data).

measurement errors could mean that the real $\tilde{e}$ is scale-independent. At present, this question remains unresolved. If future research finds that $\tilde{e}$ varies as suggested by figure 12, then the observation would imply that large and small earthquakes are dynamically different.

4. Rupture processes

We have now provided an overview of the stresses that generate earthquakes along with a discussion of the measurable parameters and their interrelationships. The next step in our inquiry into why earthquakes happen is to examine the rupture process itself.

4.1. Fracture mechanics

To interpret seismological data, crack models are often used in part because the theories on cracks have been developed well. On the other hand, seismic faulting may be more intuitively viewed as sliding on a frictional surface (fault) where the physics of friction, especially stick slip, plays a key role. Seismic faulting in the Earth can be complex and we may require a mixture of crack models and sliding models, or even other models to interpret it. Despite this complexity, crack models and frictional sliding models provide a useful framework for the interpretation of earthquake processes. Here, we limit our discussion to the very basic aspects of these models.

4.1.1. An overview of the crack model. In crack mechanics, three types of crack geometries, Mode I (tensile), Mode II (longitudinal shear) and Mode III (transverse shear), are used

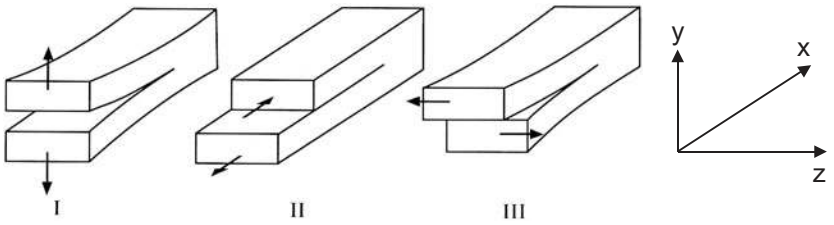

Figure 13. Modes I, II and III cracks (from Lawn (1993)). The x, y and z axes indicate the coordinate system used in figure 14.

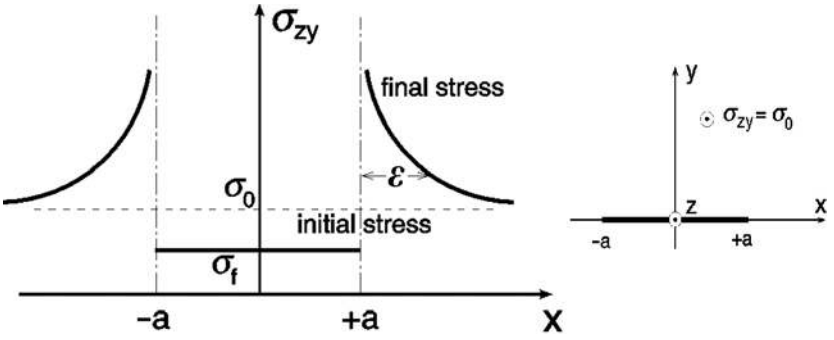

Figure 14. Stress field for a Mode III crack before and after crack formation. Dashed line is the initial stress and solid line is the final stress. The inset shows the Mode III crack geometry.

(figure 13). Although the difference between these models is important, for many problems in crack mechanics, here we mainly use the Mode III crack for illustration purposes.

Figure 14 shows the stress distribution along the plane of a Mode III crack before (dashed) and after (heavy curves) crack formation. The crack extends from $z = -\infty$ to $+\infty$ as shown in the inset. After the crack is formed, the shear stress becomes infinitely large just beyond the crack tip, and drops to the frictional stress σ_f on the crack surface.

For the coordinate system shown, the displacement w and the stress σ_{zy} are

$$w = \left(\frac{\sigma_0 - \sigma_f}{\mu}\right)(a^2 - x^2)^{1/2} \qquad x \leqslant a \tag{4.1}$$

and

$$\sigma_{zy} = (\sigma_0 - \sigma_f)\frac{x}{(x^2 - a^2)^{1/2}} + \sigma_f \qquad x \geqslant a \tag{4.2}$$

(Knopoff 1958). At a small distance ε from the crack tip, $x = a + \varepsilon$, σ_{zy} is proportional to $1/\sqrt{\varepsilon}$. Specifically, the relationship is

$$\sigma_{zy} = \frac{K}{\sqrt{2\pi}}\frac{1}{\sqrt{\varepsilon}} + \sigma_f,$$

where K is the stress intensity factor defined by

$$K \equiv \sqrt{\pi a}(\sigma_0 - \sigma_f). \tag{4.3}$$

More detailed expressions for the stress intensity factors for Modes I, II and III cracks are given in Rice (1980), Dmowska and Rice (1986) and Li (1987).

The strain energy release per unit length in z direction is (equations (3.14)–(3.16))

$$\Delta W = \frac{(\sigma_0 + \sigma_f)DS}{2} = \Delta W_0 + \sigma_f DS, \tag{4.4}$$

where

$$D = 2\bar{w} = \left(\frac{\sigma_0 - \sigma_f}{\mu}\right)\frac{1}{a}\int_{-a}^{a}(a^2 - x^2)^{1/2}\,\mathrm{d}x = \frac{\pi a}{2}\left(\frac{\sigma_0 - \sigma_f}{\mu}\right) \tag{4.5}$$

and

$$\Delta W_0 = \frac{(\sigma_0 - \sigma_f)DS}{2} = \frac{\pi a^2(\sigma_0 - \sigma_f)^2}{2\mu} \tag{4.6}$$

and $S = 2a$ and D is the average offset across the crack. In (4.4), the second term on the right-hand side (rhs) is the frictional energy and the first term, ΔW_0, is the portion of the strain energy change that is not dissipated in the frictional process.

4.1.2. Crack tip breakdown-zone. The model discussed earlier is for the static case and it provides the basic physics of dynamic crack propagation. If the stress just beyond the crack tip becomes large enough to break the material, the crack grows. In the dynamic crack propagation problem, the theory becomes complex because of the complex stress field near the crack tip and the strain energy flux into the crack tip. Here, we discuss this problem using a simple model. More rigorous and detailed discussions are by Freund (1989) and Lawn (1993).

In the simple model described in figure 14 (called the linear elastic fracture model (LEFM)), the stress near the crack tip becomes indefinitely large (solid curve in figure 14 (inset)). In the real material this does not occur. Instead, inelastic (e.g. plastic) yielding occurs, and the stress becomes finite as shown by the broken curve in figure 15(a). The finite stress at the crack tip, σ_Y, is called the yield stress. Because of this breakdown process, the stress just inside the crack does not drop to the constant frictional level σ_f abruptly. Instead it decreases gradually to σ_f over a distance l_0 as shown by the broken curve in figure 15(a). Also, slip, D, inside the crack increases gradually to the value, D_0, expected for the case without inelastic breakdown (i.e. LEFM), as shown in figure 15(b).

At a point just beyond the crack tip, the stress and slip vary as a function of time as shown in figures 15(c) and (d), respectively. Figure 15(e) shows the shear stress σ_{yz} at this point as a function of slip D, as the crack tip passes by. The stress drops from σ_Y to the constant frictional stress σ_f over a slip D_0. This behaviour in which the stress on the fault plane decreases as slip increases is called slip-weakening behaviour, and this model is often referred to as the breakdown-zone slip-weakening model. (For the development of the concept, see Dugdale (1960), Barenblatt (1962), Palmer and Rice (1973), Ida (1972), and for more detailed discussions, see Rice (1980), Li (1987).)

4.1.3. Stability and growth of a crack. Now that we have an overview of crack physics, we can consider the stability of a crack and its growth. The theory is based on Griffith's (1920) concept which was initially developed for tensional cracks (Mode I). Here, we use the basic concept, and apply it to seismological problems. More details are given by Lawn (1993).

Consider a Mode III crack with half-length, a, as discussed earlier. When a crack with half-length a is inserted in a homogeneous medium under uniform shear stress σ_0, the strain energy is released. After subtracting the energy dissipated in friction, we obtain the energy given by (4.6)

$$\Delta W_0 = \frac{(\sigma_0 - \sigma_f)DS}{2} = \frac{\pi a^2(\sigma_0 - \sigma_f)^2}{2\mu}, \tag{4.7}$$

which is available for mechanical work for crack extension.

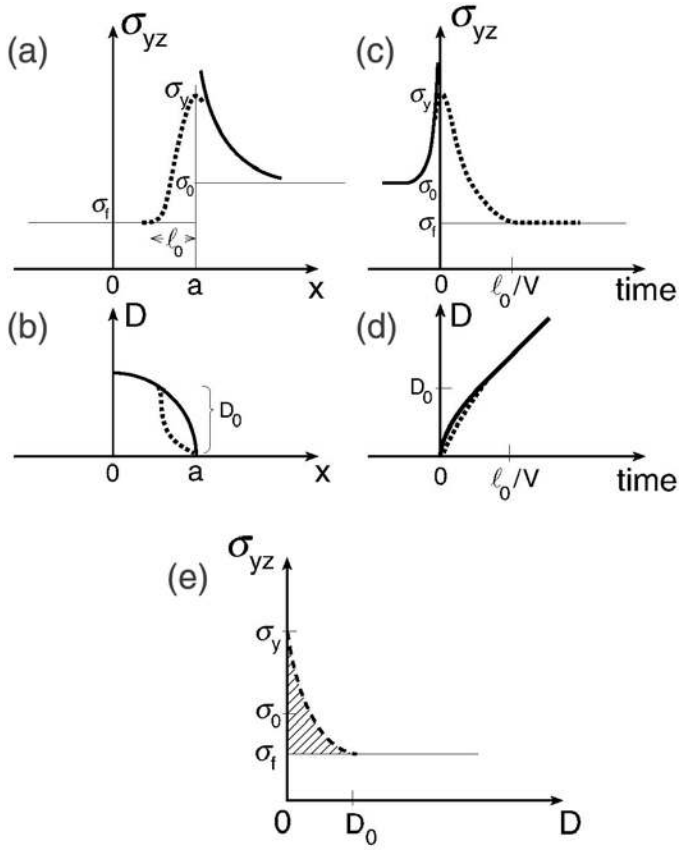

Figure 15. Breakdown-zone interpretation of slip-weakening process. (*a*) The stress field near the crack tip as a function of position. (*b*) The slip on the crack surface. (*c*) The temporal variation of stress at a point which is initially just beyond the crack tip. The time is measured from the time when the crack tip reached the point. (*d*) The temporal variation of the slip at a point which is initially just beyond the crack tip. The time is measured from the instant when the crack tip reached the point. (*e*) The shear stress at point which is initially just beyond the crack tip. As the crack tip extends past this point, the stress drops from σ_Y to σ_f gradually. In all figures, the solid curves are for the LEFM model. The broken curves indicate the deviation from the LEFM case when yielding occurs.

Now consider a virtual extension of crack by δa. Then the strain energy that would be released due to the virtual extension δa is, from (4.7)

$$\delta(\Delta W_0) = \frac{\partial \Delta W_0}{\partial a}\delta a = \frac{\pi a(\sigma_0 - \sigma_f)^2}{\mu}\delta a = 2G^*\delta a, \tag{4.8}$$

where

$$G^* = \frac{\pi a(\sigma_0 - \sigma_f)^2}{2\mu} = \frac{K^2}{2\mu}, \tag{4.9}$$

where K is the stress intensity factor defined by (4.3). G^* is called the static energy release rate or crack extension force. The name is a little confusing because here 'rate' means per unit area rather than unit time. The unit of G^* is energy per area. The factor 2 on the rhs of (4.8) arises because the crack extends at both ends. Note that G^* is not a constant, but increases as the crack size increases.

Static crack. In case of a static (or quasi-static) crack, for the crack to be stable at half-length a, this energy must be equal to twice the surface energy of the material near the crack tip. That is,

$$G^* = G_c^* \equiv 2\gamma, \tag{4.10}$$

where γ is the surface energy per unit area which is necessary to create a new crack surface. The factor 2 in (4.10) arises because surface energy is defined for each side of the crack. If G^* given by (4.9) is larger than G_c^*, the crack will grow. G_c^* is called the critical specific fracture energy. The stress intensity factor at this state, K_c, is called the fracture toughness (or critical stress intensity factor), which is related to G_c^* by equation (4.9), i.e.

$$G_c^* = \frac{K_c^2}{2\mu}. \tag{4.11}$$

Thus, the stability of a crack can be discussed either in term of the critical specific fracture energy, G_c^*, or the critical stress intensity factor, K_c.

In seismic faulting, we often generalize γ to include more surface area (e.g. damaged zones) than just the normal area of crack extension, as is done in the Griffith theory.

K and G^* are the important parameters in crack theory. The expressions for K are independent of the mode of crack. The expression for G^* is the same for Modes I and II, but is slightly different for Mode III, but considering the gross approximations used in seismological applications, the differences are not important.

Dynamic crack. When $G^* > G_c^*$ the crack propagates dynamically and some energy is radiated out of the system as seismic waves. The total energy available for work from equation (4.7) is divided between the virtual crack extension and the radiated energy. If we denote the radiated energy by E_R, the energy equation for virtual crack extension is no longer given by (4.8). Instead,

$$\delta(\Delta W_0) - \delta(E_R) = 2G\delta a, \tag{4.12}$$

from which

$$G = G^* - \frac{1}{2}\frac{\partial E_R}{\partial a}, \tag{4.13}$$

where G is called the dynamic energy release rate. Then, the crack extension is governed by

$$G = 2\gamma \tag{4.14}$$

instead of (4.10).

Rupture speed. The ratio of the dynamic energy release rate, G, to the static energy release rate, G^*, is given by a function of rupture speed $V = \mathrm{d}a/\mathrm{d}t$. Kostrov (1966), Eshelby (1969) and Freund (1972) showed that the energy release rate, G, for a crack growing at a rupture speed V is given approximately by

$$G = G^* g(V), \tag{4.15}$$

where $g(V)$ is a universal function of V.

For a Mode I (tensile) crack (Freund 1972)

$$g(V) = 1 - \frac{V}{c_R}, \tag{4.16}$$

where c_R is the Rayleigh-wave speed.

For a Mode II (longitudinal shear) crack (Fossum and Freund 1975)

$$g(V) = \frac{1 - V/c_R}{\sqrt{1 - V/\beta}}, \tag{4.17}$$

where β is the shear-wave speed.

For a Mode III (transverse shear) crack (Kostrov 1966, Eshelby 1969).

$$g(V) = \sqrt{\frac{1 - (V/\beta)}{1 + (V/\beta)}}. \tag{4.18}$$

The derivation of $g(V)$ given above is actually complicated, but the classical Mott's (1948) theory is useful for understanding the basic physics. In Mott's theory, the radiated energy E_R is equated to the kinetic energy in the medium during rupture propagation and scales as

$$E_R \propto a^2 \dot{u}^2 \propto a^2 \left(\frac{\partial u}{\partial a}\right)^2 \dot{a}^2 \propto a^2 \left(\frac{\partial u}{\partial a}\right)^2 V^2,$$

where u is the displacement. Because the strain energy, ΔW_0, also scales as a^2 (4.7), E_R can be written as

$$E_R = \frac{1}{B^2}\left(\frac{V}{\beta}\right)^2 \Delta W_0, \tag{4.19}$$

where B is a constant of the order of 1 (Lawn 1993, chapter 4; Mott 1948, Marder and Fineberg 1996). Then, including the kinetic energy, the equation corresponding to (4.12) can be written as,

$$\delta(\Delta W_0) - \delta E_R = \left[1 - \frac{1}{B^2}\left(\frac{V}{\beta}\right)^2\right]\delta(\Delta W_0) = 2G\delta a \tag{4.20}$$

from which

$$G = G^* g(V), \tag{4.21}$$

where

$$g(V) = 1 - \frac{V^2}{(B\beta)^2}. \tag{4.22}$$

Equation (4.22) has a similar form to relativistic contraction as used to calculate the electromagnetic field around a particle as it approaches the speed of light. Like the electromagnetic case, acoustic waves also experience a relativistic effect because information can only propagate through a finite distance in a finite time.

The equation for dynamic crack extension is given from (4.14) as

$$G^* g(V) = 2\gamma. \tag{4.23}$$

In the limit of the rupture speed approaching the shear sound speed β, no energy is dissipated mechanically for a Mode III crack and all the energy is radiated in elastic waves. Modes I and II cracks display the same phenomenon at different limiting velocities. In the limit of very small rupture speed, the relativistic contraction is irrelevant and $g(V)$ approaches unity.

4.2. Frictional sliding

Fault rupture can also be modelled as a frictional process. As two surfaces slide past each other along a pre-existing fault, the dynamics can be dominated by the surface forces between the two sides. Below, we quantify this frictional interaction in a way that is parallel to the crack theory so that we can combine the two formulations in section 4.3.

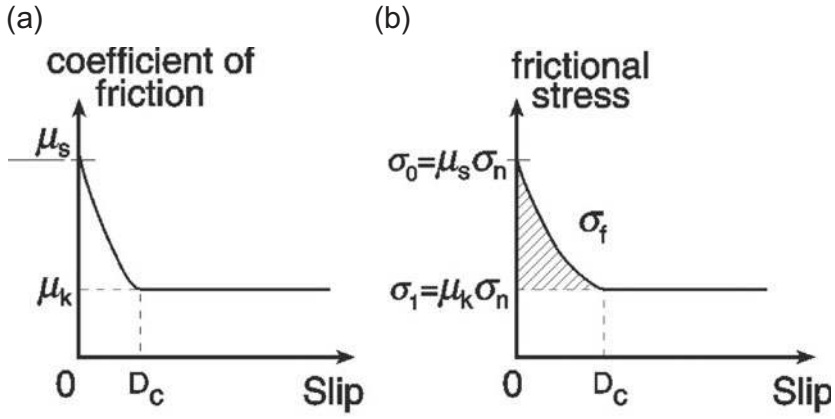

Figure 16. Static and kinetic friction. (*a*) The coefficient of friction. In the simple model, μ drops from μ_s to μ_k instantly, but in general, it drops to μ_k after a slip D_c. (*b*) The frictional stress. D_c is the critical slip. σ_f is the frictional stress.

4.2.1. Static and kinetic friction. In the classical theory of friction, the coefficient of static friction μ_s and the coefficient of kinetic friction μ_k are the most fundamental parameters (note that μ is used for coefficient of friction in this section rather than rigidity). If $\mu_k < \mu_s$, instability can occur. In any physical system, static friction cannot drop to kinetic friction instantly. A slip, D_c, is required before static friction drops to kinetic friction, and steady sliding begins (figure 16). The slip, D_c, is called the critical slip, and is a key parameter in frictional sliding models. In this case the friction σ_f is a function of slip.

The hatched area in figure 16(*b*) indicates the extra energy per unit area expended in the system compared with the case in which the friction instantly drops to the final stress, σ_1. D_c in the frictional sliding model is often equated to D_0 of the critical slip of the slip-weakening crack model (cf figure 15(*e*))

4.2.2. Rate- and state-dependent friction. The simple behaviour shown above can be generalized by a rate- and state-dependent friction model. Dieterich and his collaborators introduced the following friction law from experiments on many different materials (e.g. Dieterich (1979), Scholz (2002)). According to this law, the coefficient of friction μ is given by

$$\mu = \mu_0' + A \ln \dot{\delta} + B \ln \theta, \tag{4.24}$$

where $\dot{\delta}$ is the sliding speed, θ is a state variable that accounts for the history of sliding and μ_0', A and B are constants. Specifically, θ is governed by the following differential equation:

$$\dot{\theta} = 1 - \frac{\theta \dot{\delta}}{D_c}. \tag{4.25}$$

Although (4.24) is empirically derived, it still can be related in general terms to simple physics. The second term $A \ln \dot{\delta}$ represents a resistance similar to viscosity generated by deforming small irregularities on the sliding surface, or asperities. As they are deformed more quickly, they have a greater resisting stress. The third term $B \ln \theta$ describes the chemical adhesion between surfaces that is assumed to increase with contact time. If $\dot{\delta} = 0$, then $\dot{\theta}$ increases linearly with time (4.25), i.e. the state variable is simply the time of contact.

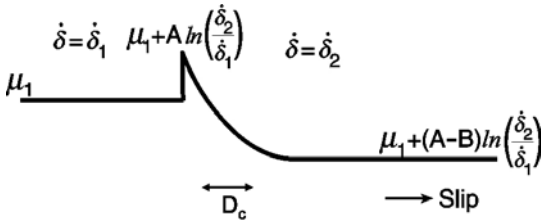

Figure 17. Change in friction due to a sudden increase in sliding speed according to the rate- and state-dependent friction.

Alternative forms of the evolution law for the state variable depend explicitly on slip rather than hold time (e.g. Ruina (1983), Linker and Dieterich (1992)). Recent experimental works favour the form in (4.25) (Beeler *et al* 1994).

If $\dot{\delta}$ = constant, then, from (4.25), θ is given by

$$\theta = \frac{D_c}{\dot{\delta}} + \left(\theta_0 - \frac{D_c}{\dot{\delta}}\right) \exp\left(-\frac{\dot{\delta}(t - t_0)}{D_c}\right), \tag{4.26}$$

where θ_0 is the value of θ at $t = t_0$.

For illustration purposes, consider a case in which $\dot{\delta}$ increases from $\dot{\delta}_1$ to $\dot{\delta}_2$ stepwise at $t = t_1$. Before $t = t_1$ we assume that sliding is in steady state at $\dot{\delta} = \dot{\delta}_1$. Then $\theta = D_c/\dot{\delta}_1$, and, from (4.24), we obtain μ for $t < t_1$ as

$$\mu_1 = \mu_0' + A \ln \dot{\delta}_1 + B \ln\left(\frac{D_c}{\dot{\delta}_1}\right) \qquad \text{for } t < t_1. \tag{4.27}$$

For $t \geqslant t_1$,

$$\theta = \frac{D_c}{\dot{\delta}_2} + \left(\frac{D_c}{\dot{\delta}_1} - \frac{D_c}{\dot{\delta}_2}\right) \exp\left[-\frac{\dot{\delta}_2(t - t_1)}{D_c}\right]$$

and

$$\mu = \mu_1 + A \ln\left(\frac{\dot{\delta}_2}{\dot{\delta}_1}\right) + B \ln\left[\frac{\dot{\delta}_1}{\dot{\delta}_2} + \left(1 - \frac{\dot{\delta}_1}{\dot{\delta}_2}\right) \exp\left(-\frac{\dot{\delta}_2}{D_c}(t - t_1)\right)\right] \qquad \text{for } t \geqslant t_1. \tag{4.28}$$

As $t \to \infty$, $\mu = \mu_2$ where

$$\mu_2 = \mu_1 + (A - B) \ln\left(\frac{\dot{\delta}_2}{\dot{\delta}_1}\right). \tag{4.29}$$

Figure 17 shows μ as a function of slip.

The cases with $(A - B) < 0$ and $(A - B) > 0$ represent velocity weakening (generally unstable) and velocity strengthening (generally stable), respectively. If $(A - B) < 0$, then the friction initially increases, but eventually drops to $\mu_2 = \mu_1 + (A - B)\ln(\dot{\delta}_2/\dot{\delta}_1)$ from μ_1. The constant D_c is a scaling parameter which determines the amount of slip over which the friction drops substantially. For example, let D_c' be the slip over which the friction drops by $(A - B)\ln(\dot{\delta}_2/\dot{\delta}_1)/e$. D_c' is proportional to D_c but it also depends on the velocity ratio $\dot{\delta}_2/\dot{\delta}_1$, and can be interpreted as the critical slip in the simple friction law shown in figure 16. For a small ratio of $\dot{\delta}_2/\dot{\delta}_1$, $D_c' \approx D_c$, but for a very large ratio of $\dot{\delta}_2/\dot{\delta}_1$, $D_c' \approx$ (10–20)D_c. Here, we do not distinguish D_c' and D_c, but if $\dot{\delta}_2/\dot{\delta}_1$ is very large, D_c and D_c' must be distinguished.

The behaviour shown in figure 17 is what was observed experimentally for many different kinds of materials (Dieterich 1979), and is the basis of the rate- and state-dependent friction law.

4.3. The link between the crack model and the friction model

Crack and frictional sliding models are frequently used in seismology to explain various aspects of earthquake phenomena. Sometimes the terminology of crack mechanics is used and at other times the terminology of friction mechanics is used. We need to link these two models to understand the way these models are used in seismology.

The fracture energy, G_c, in crack theory is the energy needed to create new crack surfaces near the crack tip. Thus, the system must expend the threshold fracture energy G_c before the crack can extend. In contrast, in frictional sliding model, D_c, is introduced as a critical slip before rapid sliding begins at a constant friction. The final value of the frictional stress σ_f is equal to σ_1. If the frictional stress varies more or less linearly as shown in figure 16, the energy spent in the system before this happens can be approximately written as (hatched area in figure 16(*b*))

$$\tfrac{1}{2}(\sigma_0 - \sigma_1)D_c. \tag{4.30}$$

Thus, if we are to link a crack model to a friction model, we can equate equation (4.30) to G_c, i.e.

$$G_c = \tfrac{1}{2}(\sigma_0 - \sigma_1)D_c. \tag{4.31}$$

Past the initial breakdown-zone, σ_f is the same in the crack and friction models.

Direct determination of D_c. With the recent availability of high-quality seismograms at short distances, it is now possible to determine a bound on D_c directly from seismograms (Ide and Takeo 1997). With inversion of seismic data, the slip at a point on the fault plane, $u(t)$, can be determined as a function of time. Then solving the equation of elastodynamics on the fault plane, we can determine the shear stress, $\sigma(t)$ as a function of time. Eliminating t from $u(t)$ and $\sigma(t)$ leads to the slip dependence of stress $\sigma(u)$ from which D_c can be estimated. For the 1995 Kobe earthquake, Ide and Takeo (1997) found D_c to be of the order of 0.5 m in the deeper part of the fault plane. However, in the process of inversion low-pass filtering is applied to the data, which tends to smooth the resulting $\sigma(u)$ versus u relationship. Thus, the critical slip, D_c, thus determined is an upper bound. Mikumo *et al* (2003) developed a method to estimate D_c directly from slip-velocity records using elastodynamic modelling. With this method, D_c for large earthquakes is also estimated to be of the order of 1 m.

The values of D_c determined by laboratory friction experiments (equation (4.28)) are approximately 5 orders of magnitude less than the upper bound derived from seismic studies. Therefore, we conclude that either the seismically determined bound on D_c is so extreme that comparison with the laboratory values is not meaningful, or the slip-weakening process at large slips is different from that of laboratory friction process. For example, Marone and Kilgore (1933) suggest that D_c is controlled by the thickness of fault gouge layers as well as the surface roughness. Ohnaka and Shen (1999) proposed a scaling relation between D_c and the wavelength of the surface roughness. Alternative slip-weakening processes include thermal pressurization, hydrodynamic lubrication (as discussed in section 4.5), plastic deformation and micro-fracturing in the crust surrounding the fault.

4.4. Rupture energy budget

Since the crack and frictional processes are linked through the fracture energy G_c, we can relate the macroscopically observable energy budget to the microscopic processes in a surprisingly general way. Any constraint on fracture energy obtained from the energy budget will provide a strong bound on all microscopic rupture processes.

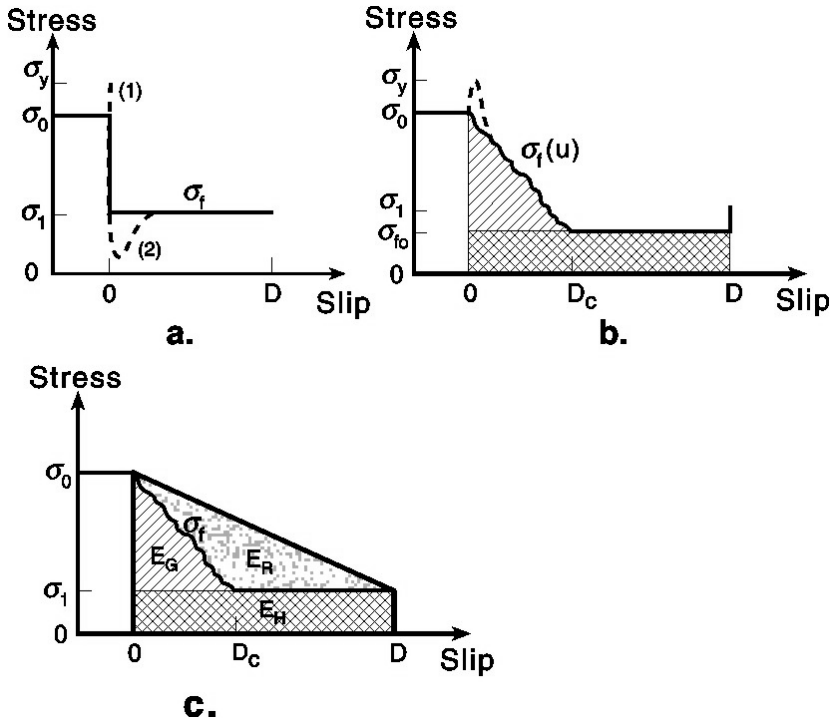

Figure 18. Illustration of simple stress release patterns during faulting. (*a*) ——: simple case of immediate stress drop. - - - -: general case without slip-weakening. (*b*) Slip-weakening model: hatched and cross-hatched areas indicate the fracture energy and frictional energy loss, respectively. (*c*) The energy budget: hatched, cross-hatched and dotted areas indicate the fracture energy, thermal (frictional) energy and radiated energy, in that order. All the figures are shown for unit area of the fault plane.

An earthquake is viewed as a stress release process on a fault surface S . The solid lines in figure 18(*a*) show the simplest case. At the initiation of an earthquake, the initial (before an earthquake) shear stress on the fault plane σ_0 drops to a constant dynamic friction σ_f, and stays there, i.e. $\sigma_f = \sigma_1$. If the condition for instability is satisfied (Brace and Byerlee (1966), Scholz (2002), also section 6.1.1), rapid fault slip motion begins and eventually stops. At the end, the stress on the fault plane is σ_1 (final stress) and the average slip (offset) is D. The difference $\Delta\sigma_s = \sigma_0 - \sigma_1$ is the static stress drop. During this process, the potential energy (strain energy plus gravitational energy) of the system, W_0, drops to $W_1 = W_0 - \Delta W$ where ΔW is the strain energy drop, and the seismic wave is radiated carrying an energy E_R. From equation (3.14),

$$\Delta W = \bar{\sigma} D S, \tag{4.32}$$

where $\bar{\sigma} = (\sigma_0 + \sigma_1)/2$ is the average stress during faulting (section 3.1.4). Graphically, ΔW (for unit area) is given by the trapezoidal area shown in figure 18(*c*).

The variation of stress during faulting can be more complex than shown by the solid lines in figure 18(*a*). For example, the stress may increase to the yield stress σ_Y in the beginning of the slip motion (curve (1) in figure 18(*a*)) because of loading caused by the advancing rupture (figure 15(*e*)), or of a specific friction law such as the rate- and state-dependent friction law (Dieterich 1979) (figure 17). In fact, some seismological inversion studies have shown this

increase (Quin 1990, Miyatake 1992, Mikumo and Miyatake 1993, Beroza and Mikumo 1996, Bouchon 1997, Ide and Takeo 1997). The stress difference, $\sigma_Y - \sigma_0$, is called the strength excess. However, the amount of slip during this high stress stage is small so that little energy is involved. Thus, we will not include it in our simple energy budget.

Also, the friction may not be constant during faulting. For instance, the friction may drop drastically in the beginning and later resumes a somewhat larger value (curve (2) in figure 18(*a*)), or it may decrease gradually to a constant level (figure 18(*b*)). As discussed before, this behaviour in which the stress on the fault plane gradually decreases as slip increases is often called slip-weakening (Rice 1980, Li 1987). Slip-weakening models have been considered in seismological models by Brune (1970), Kikuchi and Fukao (1988), Heaton (1990) and Kikuchi (1992).

If friction is not constant, the rupture dynamics is complicated, but for the energy budget considered here, we formulate this problem referring to a simple case shown in figure 18(*b*). The friction σ_f gradually drops to a constant value σ_{f0} until the slip becomes D_c. (For simplicity, here we assume that the final stress σ_1 is equal to σ_{f0}.) Then, we define the average friction $\bar{\sigma}_f$ by

$$\bar{\sigma}_f = \frac{1}{D}\int_0^D \sigma_f(u)\,\mathrm{d}u, \tag{4.33}$$

where u is the slip (offset) on the fault plane. Then, the total energy dissipation is given by

$$S\int_0^D \sigma_f(u)\,\mathrm{d}u = \bar{\sigma}_f DS. \tag{4.34}$$

Figure 18(*c*) shows the partition of energy. The area under the trapezoid outlined by the heavy lines represents the total potential energy change, ΔW. The area under the curve labelled as σ_f is the total dissipated energy. Then, the radiated energy, E_R, is the dotted area. Thus,

$$E_R = \Delta W - \bar{\sigma}_f DS. \tag{4.35}$$

As we discussed earlier, if we use the slip-weakening model, the hatched area in figure 18(*c*) is the fracture energy, E_G. Then, the total dissipated energy $\bar{\sigma}_f DS$ can be divided into E_G, and the frictional energy, E_H, represented by the cross-hatched area in figure 18(*c*). We should note that this partition is model dependent; nevertheless it is based on the breakdown-zone interpretation of the slip-weakening behaviour and is useful for interpretation of the energy budget.

From figure 18(*c*), we obtain

$$E_R = \frac{\sigma_0 - \sigma_1}{2}DS - E_G = \frac{\Delta\sigma_s}{2}DS - E_G = \frac{\Delta\sigma_s}{2\mu}M_0 - E_G, \tag{4.36}$$

where $M_0 = \mu DS$ is the seismic moment.

The ratio,

$$\eta_R = \frac{E_R}{E_R + E_G} \tag{4.37}$$

is called the radiation efficiency and is an important parameter which determines the dynamic character of an earthquake (Husseini 1977). The radiation efficiency, η_R, is different from the seismic efficiency, η, which is given by $E_R/\Delta W$. As discussed in section 3.1.4, ΔW cannot be determined directly by seismological methods and the seismic efficiency is difficult to determine. Because $\Delta W \geqslant E_R + E_G$, $\eta_R \geqslant \eta$. If $\eta_R \approx 1$, the breakdown zone is unimportant and failure occurs primarily in the steady-state regime regardless of whether it is crack-like or friction-dominated. On the other hand, if $\eta_R \ll 1$, the microscopic breakdown process is dominating the dynamics.

By combining (4.36) and (4.37), we obtain a relation between η_R and observable seismological parameters:

$$\eta_R = \frac{2\mu}{\Delta\sigma_s}\tilde{e}, \tag{4.38}$$

where

$$\tilde{e} = \frac{E_R}{M_0} \tag{4.39}$$

is the radiated energy scaled by the seismic moment, i.e. scaled energy. As shown in figure 12, $\tilde{e}$ is always less than 10^{-4}, and values of $\tilde{e}$ for small earthquakes are often 1–2 orders of magnitude less than those for large earthquakes. Whether the trend shown in figure 12 is real or not is currently debated (Ide and Beroza 2001). If it is real, for typical values of static stress drops, 1–10 MPa, and a shear modulus, 3×10^4 MPa, η_R for small earthquakes must be significantly less than unity. The small values of η_R for small earthquakes motivate us to examine the micro-mechanisms of slip-weakening and breakdown.

4.5. Fault-zone processes: melting, fluid pressurization and lubrication

Motivated by the importance of the non-elastic slip-weakening processes in the energy balance, we now turn to the micromechanics of rupture beyond solid friction and crack models.

Melting. One of the first such special mechanisms recognized was frictionally-induced melting. As first suggested by Jeffreys (1942), frictional dissipation may be high enough, early in the rupture, to melt the wallrock. The silicate melt then reduces the friction for the remainder of the earthquake. Studies by McKenzie and Brune (1972), Richards (1976) and Cardwell *et al* (1978) quantitatively confirmed the potential importance of frictional heating during faulting.

Here, we consider a gross thermal budget during faulting under a frictional stress σ_f. Let S and D be the fault area and the displacement offset, respectively. Then the total heat generated during faulting is $Q = \sigma_f DS$. If we assume that the heat is distributed during seismic faulting within a layer of thickness w around the rupture plane and there is negligible heat transport away over the timescale of the earthquake, the average temperature rise ΔT is given by

$$\Delta T = \frac{Q}{C\rho Sw} = \frac{\sigma_f D}{C\rho w}, \tag{4.40}$$

where C is the specific heat and ρ is the density. In general D increases with the earthquake magnitude, M_w. Using the scaling relations given in section 3.2.1, we can relate D to M_w (with the static stress drop $\Delta\sigma_s$ as a parameter), and compute ΔT as a function of magnitude with three parameters, σ_f, w and the static stress drop $\Delta\sigma_s$. Figure 19 shows ΔT for the case with $w = 1$ cm as a function of M_w.

If σ_f is comparable to $\Delta\sigma_s$ about 10 MPa, the effect of shear heating is significant. If the thermal energy is contained within a few centimetres around the slip plane during seismic slip, the temperature can easily rise by 100–1000°C during a moderate-sized earthquake.

Thermal fluid pressurization. Geological outcrops of faults suggest that many faults are filled with aqueous fluids or a viscous mixture of gouge and water when active. If this is true, then another set of mechanisms are at work during rupture. One possibility is thermal pressurization of the fluid. This concept was introduced to seismology by Sibson (1973), and analysed in great detail by Lachenbruch (1980), Mase and Smith (1985, 1987) and Andrews (2002). Under the

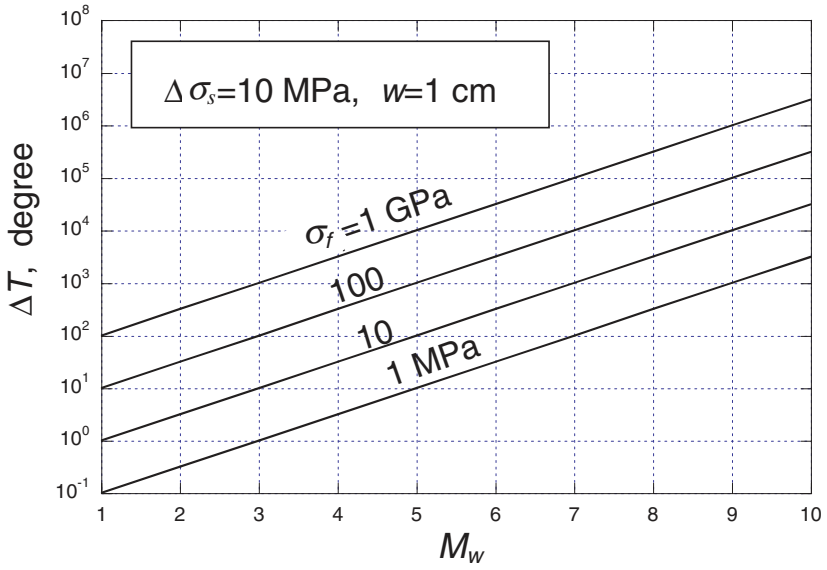

Figure 19. Predicted temperature rise, ΔT, in a fault zone as a function of magnitude, $M_{\rm w}$, with the frictional stress, $\sigma_{\rm f}$, as a parameter. The static stress drop $\Delta\sigma_{\rm s}$ is assumed to be 10 MPa.

pressure–temperature conditions at the seismogenic depths, the thermal expansivity of water is of the order of $10^{-3}\,^{\circ}\mathrm{C}^{-1}$, and significant increase in pore pressure with temperature could occur. If the fluid does not escape (small permeability) and the surrounding rock is not compressible, the pressure increase would be of the order of 1 MPa per $^{\circ}$C (Lachenbruch 1980). In actual fault zones, permeability and compressibility vary and the pressure increase may be less. The most important parameter controlling the pressure change is the permeability. The analysis of Lachenbruch (1980) and Mase and Smith (1985, 1987) suggests that if permeability is less than 10^{-18} m^2, fluid pressurization is most likely to occur with a temperature rise of less than 200°C, and the friction will drop significantly. Permeability in the crust varies over a very wide range of more than a factor of 10^{10}. Although the distribution of permeability can be complex, these studies suggest that fluid pressurization can play an important role, at least locally, in reducing friction. A modest ΔT of 100–200°C would likely increase the pore pressure enough to significantly reduce friction. Figure 19 shows that this moderate temperature increase can occur even for intermediate-sized earthquakes ($M_{\rm w} = 3$–5).

The key question is: what is the thickness w of the fault slip zone? Geologists have examined many old fault zones which were formed at depths and were brought to the surface by long-term uplift (i.e. exhumed faults). Some fault zones have a very narrow (about 1 mm) distinct slip zone where fault slips seem to have occurred repeatedly. According to Chester and Chester (1998), the internal structure of the Punchbowl fault, California, implies that earthquake ruptures were not only confined to a layer of finely shattered rock, but also largely localized to a thin prominent fracture surface. They suggest that mechanisms that are consistent with the extreme localization of slip, such as thermal pressurization of pore fluids, are most compatible with their observations. In other cases, several narrow slip zones were found but evidence shows that each slip zone represents a distinct slip event (i.e. an earthquake). Thus, geological evidence suggests a narrow slip zone, at least for some faults, but this question remains debatable (Sibson 2003).

Lubrication. If a fault zone is narrow and slightly rough, and if the material in the fault zone behaves as a viscous fluid, it is also possible that elastohydrodynamic lubrication plays an important role in reducing friction for large events (Brodsky and Kanamori 2001). As in a ball bearing, the transverse shear gradients in the fluid are balanced by the longitudinal pressure gradients and the pressure increases on the leading edge of irregularities in the fault surface. An interesting consequence of this is that as the slip and slip velocity increase, the hydrodynamic pressure within a narrow zone becomes large enough to smooth out the irregularities on the fault surface by elastic deformation, thereby suppressing high-frequency ground motion caused by the fault surfaces rubbing against each other. During the 1999 Chi-Chi, Taiwan, earthquake, the observed ground-motion near the northern end of the fault was extremely large ($>3.5\,\mathrm{m\,s^{-1}}$, the largest ever recorded), but short period acceleration was not particularly strong and the shaking damage was not the worst (Ma *et al* 1999). This could be a manifestation of the high-speed lubrication effects (Ma *et al* 2003). However, since this is the only earthquake for which such large slip and slip velocity were instrumentally observed, whether this is indeed a general behaviour or not is yet to be seen.

Any of these dynamic weakening mechanisms can explain the lack of elevated heat flow over seismogenic faults (section 2.2). There is some evidence that the heat flow is slightly elevated over the section of the San Andreas fault that slips gradually with no large earthquakes (Sass *et al* 1997). A high heat flow over the aseismic section would be consistent with dynamic weakening reducing the frictional dissipation where there are earthquakes.

Since a fault zone is probably complex and heterogeneous in stress, fluid content, permeability, porosity and compressibility, no single process is likely to dominate. In other words, we do not necessarily expect a single continuous layer of melting and pressurization; we envision, instead, a fault zone that consists of many regions where different mechanisms are responsible for slip at different stress levels, producing complex rupture patterns as observed.

In these discussions, the thickness of fault slip zones is the key parameter for understanding fault dynamics. Of course, whether lubrication occurs or not depends on many factors such as the effective permeability in the fault zone, compressibility of fault rocks and the viscosity of melts; but in view of the large slip and slip velocity associated with seismic faulting, a significant drop in friction is likely to occur if the slip zone is narrow.

4.6. Linking processes to the seismic data

4.6.1. The interpretation of macroscopic seismological parameters

Radiation efficiency. As we discussed in section 4.4, in the breakdown-zone interpretation of the slip-weakening model, the energy defined by the cross-hatched area in figure 18(*c*) is interpreted as frictional thermal energy, E_{H}, and is subtracted from the total potential energy; it does not directly control the dynamics of earthquake rupture. In contrast, the fracture energy, E_{G}, represents the mechanical energy loss during faulting, and controls the fault dynamics in a fundamental way. Thus, the determination of fracture energy for earthquakes is critically important for understanding the dynamics of faulting. Since the radiation efficiency, η_{R}, is directly related to E_{G} by (4.37), first we describe how we can determine η_{R} from macroscopic seismic parameters (Venkataraman and Kanamori 2004).

As shown by equation (4.38), we can estimate the radiation efficiency, η_{R}, using the three macroscopic seismological parameters, M_0, E_{R} and $\Delta\sigma_{\mathrm{s}}$. If $\eta_{\mathrm{R}} = 1$, no energy is dissipated mechanically and the potential energy, after heat loss has been subtracted, is radiated as seismic waves and the earthquake is considered a very 'brittle' event. In contrast, if $\eta_{\mathrm{R}} = 0$, the event is quasi-static and no energy is radiated, even if the static stress drop is very large.

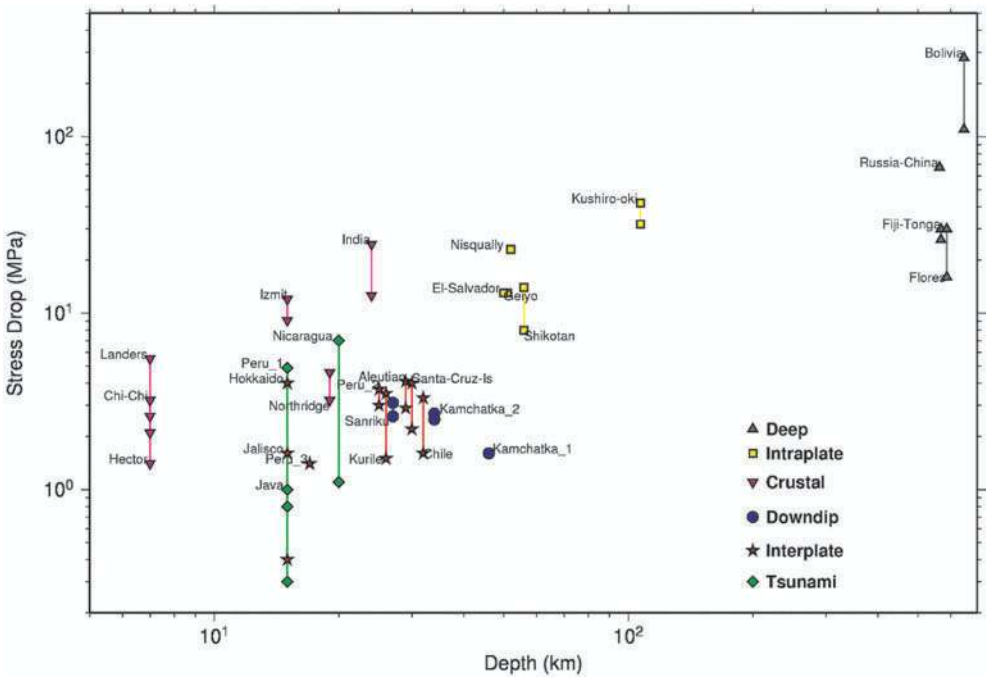

Figure 20. Static stress drop plotted as a function of depth for the different types of earthquakes. Deep: deep earthquakes; Intraplate: earthquakes which occur within the lithospheric plate; Crustal: earthquakes which occur within continental crusts; Downdip and Interplate: earthquakes which occur on the subduction-zone plate boundary; Tsunami: earthquake with a slow deformation at the source which generates tsunamis disproportionately large for its magnitude (figure taken from Venkataraman and Kanamori (2004)).

For many large earthquakes, the seismic moment, M_0, and the radiated energy, E_R have been determined. The determination of the static stress drop, $\Delta\sigma_s$, is a little more difficult. Although the scaling relation discussed in section 3.2.1 (figure 11) shows that most large earthquakes have comparable stress drops in the range of 1–10 MPa, we need to determine the stress drops for individual earthquakes for this purpose. Figure 20 shows the estimates for large earthquakes.

For shallow earthquakes, $\Delta\sigma_s$ is in the range of 1–10 MPa, as discussed in section 3.2. We can see a general trend of $\Delta\sigma_s$ increasing with depth. For the deepest earthquakes, $\Delta\sigma_s$ is in the range of 20–200 MPa, and the average is roughly 10 times larger than that for shallow earthquakes. This trend is roughly consistent with the result of Houston (2001) who found that the source duration of deep earthquakes with comparable magnitudes is systematically shorter than that of shallow earthquakes. This result, if interpreted using the assumption that the rupture speed is, on the average, similar to the S-wave speed, suggests a trend similar to that shown in figure 20.

Using the estimates of radiated energy, seismic moment and static stress drop, we can determine the radiation efficiency for all these earthquakes using (4.38) (Venkaraman and Kanamori 2004). Figure 21 shows the radiation efficiencies as a function of the magnitude, M_w. For a few earthquakes the computed η_R is larger than 1. This is probably due to the errors in the estimates of radiated energy and/or stress drops.

The radiation efficiency of most earthquakes lies between 0.25 and 1. Tsunami earthquakes (earthquakes with slow deformation at the source which generate tsunamis disproportionately

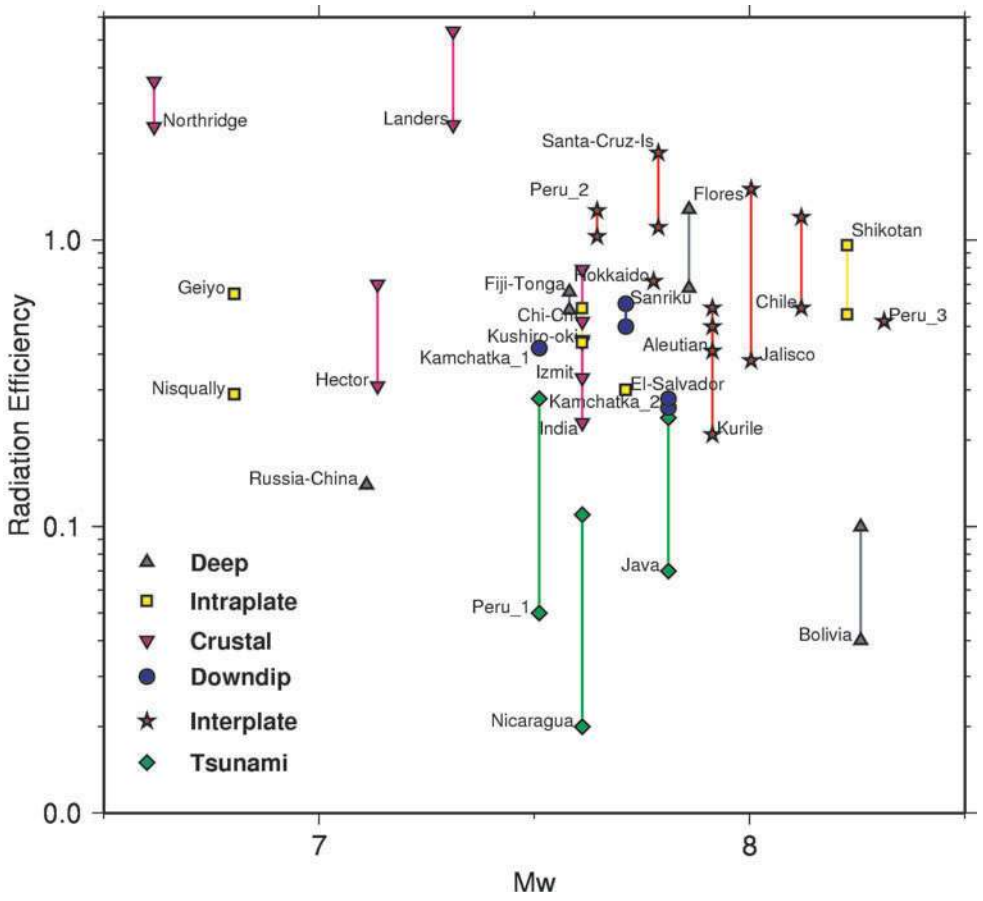

Figure 21. Radiation efficiency $\eta_R = E_R/(E_R + E_G)$ as a function of M_w. The different symbols show different types of earthquakes as described in figure 20. Most earthquakes have radiation efficiencies greater than 0.25, but tsunami earthquakes and two of the deep earthquakes (the Bolivia earthquake and the Russia–China border earthquake) have small radiation efficiencies (figure taken from Venkataraman and Kanamori (2004)).

large for its magnitude), however, have small radiation efficiencies (<0.25) and the two deep earthquakes, the 1999 Russia–China border event and the 1994 deep Bolivia earthquake, have small radiation efficiencies.

For the 1994 Bolivian earthquake ($M_w = 8.3$, depth $= 635$ km), the largest deep earthquake ever recorded, the source parameters could be determined well enough to investigate the energy budget (Kanamori *et al* 1998). The result showed that $\Delta W_0 = 1.4 \times 10^{18}$ J and $E_R = 5 \times 10^{16}$ J, which is only 3% of ΔW_0, and the difference $\Delta W_0 - E_R = 1.35 \times 10^{18}$ J, was not radiated, and must have been deposited near the focal region, probably in the form of fracture energy in addition to the frictional energy. This energy 1.35×10^{18} J is comparable to the total thermal energy released during large volcanic eruptions such as the 1980 Mount St Helens eruption. In other words, fracture and thermal energy at least comparable to that released by a large volcanic eruption must have been released in a relatively small focal region, about 50×50 km^2, within a matter of about 1 min. The mechanical part of the process, i.e. the earthquake observed as seismic waves, is only a small part of the whole process. Thus, the Bolivia earthquake should be more appropriately viewed as a thermal process rather than

a mechanical process. How much of the non-radiated energy goes to heat depends on the details of the rupture process, which is unknown. However, it is possible that a substantial part of the non-radiated energy was used to raise the temperature in the focal region significantly. The actual temperature rise, ΔT, also depends on the thickness of the fault zone, which is not known, but if it is of the order of a few centimetres, the temperature could have risen to above 10 000°C (figure 19).

The relation between radiation efficiency and rupture speed. As discussed in section 4.1.3, the energy release rate for a dynamic crack G is given as a function of the rupture speed V (equation (4.15)). Since the fracture energy E_G can be interpreted as the integral of G over the entire fault surface S, we can relate the radiation efficiency, η_R, to rupture speed, V as follows. In the simplest model discussed in section 4.1.3, we can use equations (4.7) and (4.9) to write

$$E_G = \int G \, dS = g(V) \int G^* \, dS = g(V) \Delta W_0 \tag{4.41}$$

and

$$\eta_R = \frac{E_R}{E_R + E_G} = \frac{E_R}{\Delta W_0}, \tag{4.42}$$

from which we obtain

$$\eta_R = 1 - g(V). \tag{4.43}$$

The average rupture speed is usually determined from the inversion of seismic waves and the results can be non-unique, but for large earthquakes, the estimates of rupture speed are fairly accurate. Most of these earthquakes have rupture speeds such that the ratio of rupture speed to shear wave speed (V/β) is between 0.75 and 0.95. However, the 1994 Bolivia earthquake, the 1999 Russia–China border event and the tsunami earthquakes, have small V/β, about 0.1 to 0.2. Figure 22 shows the upper and lower limit of radiation efficiencies that were determined from the energy budget plotted against the upper and lower limit of V/β obtained from the literature. The theoretical curves relating radiation efficiency to rupture speed for Modes I, II and III cracks (equations (4.16)–(4.18)) are also plotted in the same figure. To the first order, the observed data follow the theoretical curves obtained from crack theory. Since rupture speed is an independently determined quantity, this consistency of the observed relationship between η_R and V/β with the calculations from crack theory enhances the results shown in figure 21.

Summary and implications. With the three seismologically observable macroscopic parameters (seismic moment M_0, radiated energy E_R and the static stress drop $\Delta\sigma_s$), we showed that for most earthquakes, the radiation efficiency which is given by $\eta_R = E_R/(E_R + E_G)$ is larger than 0.25, which means that the amount of energy mechanically dissipated during rupture is comparable or smaller than the energy radiated as seismic waves. This conclusion seems to be supported by the independent observations of the high rupture speed V. Note that this line of reasoning poses no constraint on the energy E_H dissipated directly as heat.

For tsunami earthquakes (slow seismic events) and some deep earthquakes, the radiation efficiency is small, which means that the rupture process of these earthquakes involves more dissipative processes than the average. One interpretation is that most tsunami earthquakes involve rupture in soft deformable sediments, and a large amount of energy is used in deformation. The mechanism of large deep earthquakes is not known well, but it is likely that the rupture process in the pressure–temperature environment at large depths may involve large amounts of plastic deformation with large amounts of energy dissipation.

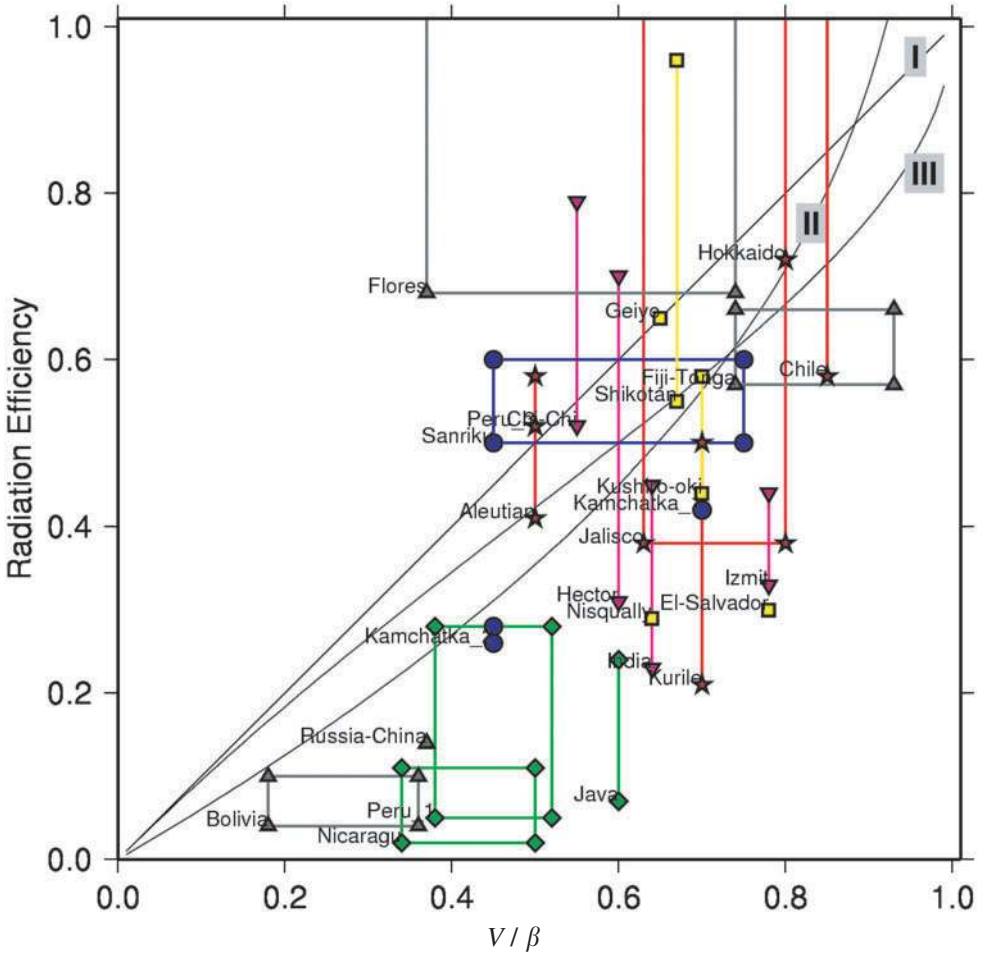

Figure 22. Radiation efficiencies determined from the radiated energy-to-moment ratios and estimates of static stress drop (equations (4.38) and (4.39)) plotted against the estimates of the ratio of rupture speed to shear wave speed obtained from literature. The symbols are the same as in figure 20; for comparison, the theoretical curves relating radiation efficiency to rupture speed for Modes I, II and III cracks have also been plotted (figure taken from Venkataraman and Kanamori (2004)).

The relatively large radiation efficiency, i.e. relatively small critical fracture energy G_c or small fracture toughness K_c (4.11), for most shallow earthquakes has an important implication for rupture growth of earthquakes. As discussed in section 4.1.3, the rupture growth is controlled by the balance between the dynamic stress intensity factor K and K_c. As a rupture grows, the length scale a increases and K increases (equation (4.3)). Thus, if K_c is small, on the average, the rupture is more likely to grow and develop into a runaway rupture. If friction decreases as the slip increases, as discussed in section 4.5, the tendency for runaway would increase because K is also proportional to $(\sigma - \sigma_f)$ (equation (4.3)). If this is the case, once an earthquake is initiated, it will be difficult to stop the rupture dynamically. To stop the rupture, some external static features such as a strength barrier or irregular fault geometry may be required. In terms of the friction model discussed in section 4.3, the small G_c or K_c means small critical slip, D_c (equation (4.31)).

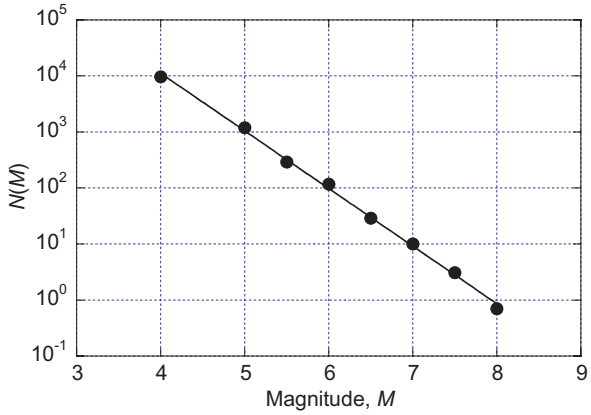

Figure 23. Magnitude–frequency relationship for earthquakes in the world for the period 1904 to 1980. $N(M)$ is the number of earthquakes per year with the magnitude $\geqslant M$. The solid line shows a slope of -1 on the semilog plot which corresponds to a b-value of 1. Note that, on the average, approximately one earthquake with $M \geqslant 8$ occurs every year. The data sources are as follows: $M \geqslant 8$, for the period 1904 to 1980 from Kanamori (1983); $M = 5.5$, 6.0, 6.5, 7.0 and 7.5, for the period from 1976 to 2000 from Ekstrom (2000); $M = 4$ and 5, for the period January 1995 to January 2000 from the catalogue of the Council of National Seismic System. For this range, the catalogue may not be complete, and N may be slightly underestimated.

At present, the accuracy of the macroscopic source parameters, especially E_R and $\Delta\sigma_s$, is not good enough to accurately estimate the fracture parameters G_c, K_c and D_c, and to draw more definitive conclusions on the rupture dynamics of earthquakes. Currently, extensive efforts are being made to improve the accuracy of determinations of the macroscopic source parameters.

5. Earthquakes as a complex system

Another possible approach to understanding why earthquakes happen is to take a broad view beyond a single event. We can study earthquakes by dealing with large groups of earthquakes statistically. The goal is to find systems that robustly reproduce the general patterns of seismicity regardless of the details of the rupture microphysics. This approach has had considerable success characterizing the types of models that will reproduce the observed magnitude–frequency relationship (i.e. Gutenberg–Richter relation) used in seismology.

The magnitude–frequency relationship (the Gutenberg–Richter relation). In general small earthquakes are more frequent than large earthquakes. This is quantitatively stated by the Gutenberg–Richter relation (Gutenberg and Richter (1941), a recent review is found in Utsu (2002).) It describes the number of earthquakes expected of each size, or magnitude, in a given area. In any area much larger than the rupture area of the largest earthquake considered, the number of earthquakes, $N(M)$, which have a magnitude greater than or equal to M is given by the relation

$$\log N(M) = a - bM, \tag{5.1}$$

where a and b are constants. Figure 23 shows that the Gutenberg–Richter relationship even applies to a seismicity catalogue encompassing the entire planet. Approximately one earthquake with $M \geqslant 8$ occurs every year somewhere in the Earth.

For most regions the value of b is nearly 1, as is the case in figure 23. This strikingly consistent observation has motivated much of the study on fault networks and self-organized criticality. (For more extensive discussions on this subject, see Main (1996), Turcotte (1997), Rundle *et al* (2000) and Turcotte and Malamud (2002).) The primary conclusions of this research are that over a wide (but finite) range of scales, fault networks are fractal. Cascades of failure result when the faults are extremely close to failure prior to any late-stage triggering. These cascades can be interpreted as an example of self-organized criticality.

The Gutenberg–Richter Law is a major tool in probabilistic hazard assessment. It allows extrapolation from the rates of small earthquakes, which we observe easily, to the likelihood of large events. Given the societal importance of the resulting hazard assessments, it is important that the empirically-derived Gutenberg–Richter Law be put on firm physical ground.

Simple models. The starting point of complexity models is the assumption that the initiation, growth and cessation of earthquake rupture are controlled by the complex interaction of fault-bounded blocks on scales as small as individual cracks and as large as continents. Because of the large number of elements involved and of the complexity of the interaction it may not be determinable exactly how different parts of the crust interact with each other, well enough to understand the earthquake process in a deterministic way. Nevertheless, some properties of earthquakes such as the magnitude–frequency relationship can be understood as manifestations of the general behaviour of complex systems.

The crux of a successful earthquake complexity model is the recreation of critical behaviour by setting up a system of elements with only local interactions being specified. A critical state is when events of all sizes can occur and their frequency distribution follows a power law. In the critical state, the local interactions can accumulate to generate long-range organization. Self-organized criticality is when a system evolves to the critical state naturally without any dependence on the initial conditions or tunable parameters (Bak *et al* 1988, Hergarten 2002). In practice, systems are only critical within a certain range of scales determined by the overall boundaries of the system. Because of the ubiquity of the Gutenberg–Richter distribution, earthquakes are thought to be self-organized critical systems. Three specific models for generating the critical state in earthquake processes are the mechanical slider-block system, the percolation model and the sand pile model.

In the slider-block system (Burridge and Knopoff (1967), figure 24(a)), many blocks are connected by a spring and the whole mass–spring system is dragged on a frictional surface. The friction between the block and the surface is governed by a simple velocity-weakening law. As the mass–spring system is dragged, some blocks slip intermittently. Most of the time, a single block slips. This slip is interpreted as a small 'earthquake', and some potential energy is released. However, occasionally slip of a block triggers slip of adjacent blocks causing a larger earthquake with a larger amount of potential energy release. If more than one block are triggered, we have a larger earthquake. After a series of events have occurred, we count the number of events $N(E)$ which released a potential energy larger than E. As shown in figure 24(b), the relation between $N(E)$ and E exhibits a power-law like behaviour, and when plotted on a log–log diagram, the relation looks like the earthquake magnitude–frequency relationship. In this slider-block model, the interaction between the blocks is described by a differential equation that involves the spring stiffness, mass and friction.

In a percolation model (Otsuka 1971), a seismic fault is modelled by a distribution of many small elements (patches). If one patch fails, then it can trigger failure of the adjacent patches with some probability. This process continues until it stops spontaneously, and the whole continuous failure corresponds to an earthquake rupture. If a patch can trigger nearby patches at s sites with a transition probability, p, then $e = ps$ is the expectancy of the number of patches

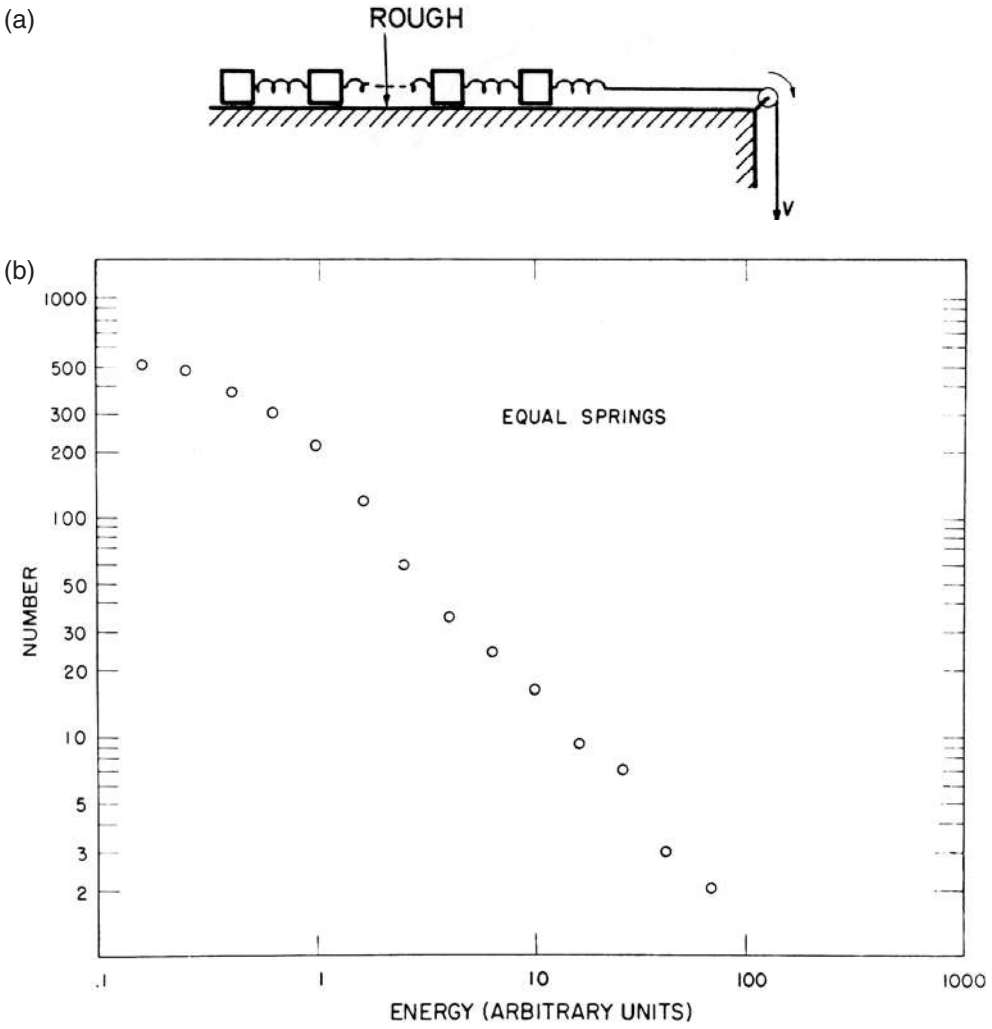

Figure 24. (*a*) Mass–spring system sliding on a frictional surface (Burridge and Knopoff 1967). (*b*) The relation between the number of events and released potential energy (Burridge and Knopoff 1967).

to be triggered at each step. If $e < 1$, then the growth will eventually stop at a certain step when a total of F patches have failed. This corresponds to a sub-critical state, i.e. the range of interactions is limited. If this whole process is repeated many times, we find a relation similar to the magnitude–frequency relationship between log F and the number of cases, n, in which at least F patches failed. Figure 25 shows the results of numerical simulations performed for $s = 10$ and three cases, $e = 0.8$, 0.9 and 0.99. When $e = 1$, the system is critical. This distribution corresponds to the magnitude–frequency distribution shown in figure 23.

In a sand pile model, a sand particle is added to a sand pile which has been maintained in a critically stable state. Usually a small number of sand particles slide off the pile and into adjacent areas as a result of the small additional stress. However, there is also a small but finite probability that a large slide can happen as a result of simply adding one particle. Bak *et al* (1988) used a set of simple rules that simulate the actual potentially complex physical

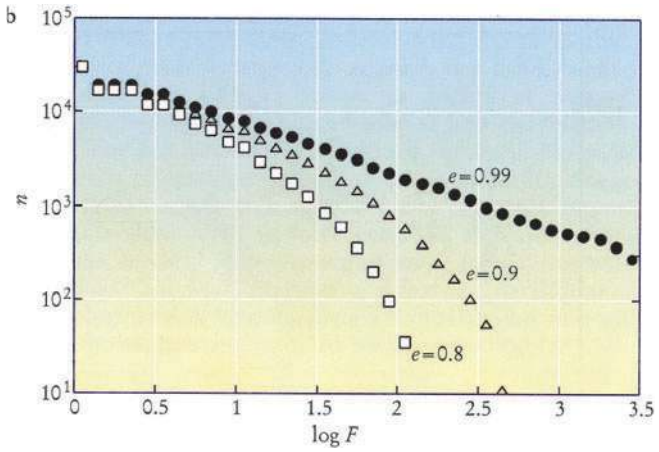

Figure 25. Magnitude–frequency relationships produced with a percolation model (figure taken from Kanamori and Brodsky (2001)).

interaction between different cells. The result of the interaction yields a relationship like the earthquake magnitude–frequency relationship (Bak and Tang 1989, Kadanoff 1991).

In all these models, what is essential is the interaction between the many elements which make up the system. In the actual earthquake process, the interaction is between different parts of a fault and between different faults. This interaction can be due to static as well as dynamic processes. The robustness of the Gutenberg–Richter result tells us that this particular feature of seismicity is insensitive to the microscopic physics controlling the failure and rupture. The details of the microphysics are only important for addressing other questions, such as the likelihood of a particular earthquake on a particular fault. The complexity models also show that earthquake interactions have inherently chaotic elements in addition to the predictable elements governed by the stress loading mechanism and fault structures. Unravelling the limits and extent of the chaos is prerequisite for determining whether or not individual earthquakes are predictable over societally useful timescales given the limited resolution of our observations of the initial conditions.

In the above, we discussed only the magnitude–frequency relationship. Another commonly observed seismicity pattern is the timing of aftershock decay which is known as Omori Law. As we will discuss later (section 6.2.4), Omori's Law can be explained with several physical models.

6. Instability and triggering

We have now covered the major pieces of the earthquake puzzle: stress in the crust, observable parameters, macroscopic observations, micromechanics and complex systems. We are now ready to use these tools to broach the key problem of earthquake triggering and instability. We will first deal with frictional instability and then approach data that has bearing on other types of initiation.

6.1. Instability

6.1.1. Stick slip and instability. Friction on a fault is not constant. Even in simple high-school formulations, friction depends on slip velocity as dynamic friction is less than static friction.

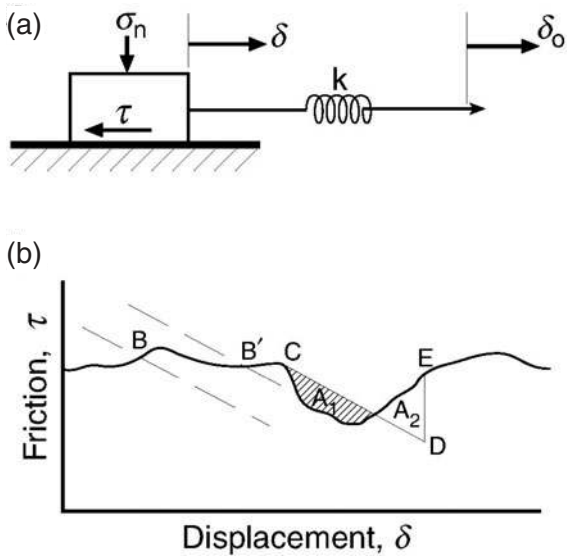

Figure 26. Stick slip model. (*a*) A spring-loaded slider block on a frictional surface. (*b*) Variation of friction as a function of displacement (——) and the spring loading force (– – –).

In realistic fault conditions, the slip-velocity dependence can be more complex and friction also varies as a function of time and slip distance. Therefore, sliding does not occur smoothly; it occurs in a stop-and-go fashion. This behaviour is generally called stick slip (Brace and Byerlee 1966).

Figure 26 shows a mechanism of stick slip (e.g. Rabinowicz (1995)) illustrated by a spring-loaded slider-block model similar to that discussed in section 5. In this figure, k, δ, δ_0, σ_n and τ are the spring constant, the displacement of the block, the displacement of the right-hand end of the spring, the normal stress and the friction between the block and the surface, in that order. Suppose we increase δ_0 by pulling the spring from the right. Then, the force balance is given by

$$\tau = k(\delta_0 - \delta) = -k\delta + k\delta_0, \tag{6.1}$$

where

$$\tau = \mu\sigma_n. \tag{6.2}$$

The solid curve in figure 26(*b*) shows the variation of τ as a function of δ. The broken line in figure 26(*b*) is the loading force exerted by the spring, given by the rhs of (6.1). The intersection, B, between the broken line and the solid curve gives the equilibrium position given by equation (6.1). As δ_0 increases, the broken line moves upward, and the intersection moves to B′. Between points B and C, the block moves over the surface smoothly. At point C, τ drops suddenly and the spring force exceeds τ, and the block moves abruptly along C–D driven by the spring force. The area A_1 is approximately equal to A_2. The block is stationary at D until the spring force reaches point E with the increase of δ_0, from where smooth motion begins again.

More precisely, at point C,

$$\left|\frac{\Delta\tau}{\Delta\delta}\right| = k \tag{6.3}$$

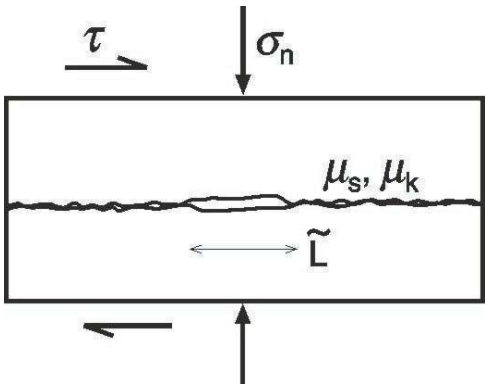

Figure 27. A schematic figure showing the nucleation length, $\tilde{L}$, on a frictional fault plane. For a rupture to grow, $\tilde{L}$ must be larger than $\tilde{L}_n$ given by (6.7).

and the stick slip instability begins past this point where

$$\left|\frac{\Delta\tau}{\Delta\delta}\right| > k. \tag{6.4}$$

If the spring constant k is large, the slope of the broken line increases, and no instability occurs, i.e. stable sliding occurs without stick-slip behaviour. Thus, the spring constant (stiffness of the system) controls the stability for a given frictional property of the surface.

Stiffness of the fault system. In the models described above, the stiffness of the spring, k, plays a key role in determining the stability. Then the question is: what is the stiffness of the crust?

Stiffness is defined by the ratio of stress to displacement, e.g. $k = \tau/\delta$. Then, if we consider a small crack with length scale $\tilde{L}$ in a nucleation zone, the stress required to cause a slip D is given by $ED/\tilde{L}$, where E is a relevant elastic modulus. Thus, the fault stiffness can be defined by

$$k_f \approx \frac{E}{\tilde{L}}. \tag{6.5}$$

6.1.2. Nucleation zone. Consider the nucleation of a slip on a frictional surface with the normal stress σ_n, the static friction coefficient, μ_s, kinetic friction coefficient, μ_k, and critical slip, D_c (figure 27).

Then from the stick-slip model (equation (6.3)), at the critical point,

$$k_f D_c = \sigma_n(\mu_s - \mu_k). \tag{6.6}$$

Then, combining this expression with the definition of k_f, we define a critical fault length scale $\tilde{L}_n$:

$$\tilde{L}_n \equiv \tilde{L} \approx \frac{E D_c}{(\mu_s - \mu_k)\sigma_n}. \tag{6.7}$$

According to the frictional instability model, $\tilde{L}_n$ in (6.7) is the nucleation length of this frictional surface. If we assume that the laboratory measurements of μ_s, μ_k and D_c on rocks are appropriate for natural faults, at typical seismogenic depths where the normal stress is $\sim$200 MPa, D_c is $\sim$10 μm, $\mu_s - \mu_k \sim 0.1$ and E is $\sim 5 \times 10^{10}$ Pa, then $\tilde{L}_n \approx 3$ cm. For a more sophisticated frictional model, the nucleation zone can be as large as 1 m (e.g. Lapusta and Rice (2003)). For the latter values, the strain $D_c/\tilde{L}_n$ in the nucleation zone prior to the

earthquake is of the order 10^{-5}. This relatively large strain could potentially be observed up to ~80 m away on modern strain metres. The major impediment to testing such a prediction method is the very small size of the proposed nucleation zone. A geophysicist would have to be very lucky to choose to put an instrument within 80 m of the right 1 m^2 patch of the 15 000 km^2 San Andreas fault!

The nucleation length model in (6.7) suggests that the smallest possible earthquakes have magnitudes $M_w = -1$ (from equation (3.7). $M_w = -1$ corresponds to $M_0 = 4 \times 10^7$ N m). Smaller earthquakes have been observed, although they are difficult to detect. Dense networks designed to capture extremely small earthquakes could help confirm or refute the nucleation length model by determining whether any lower bounds on earthquake size exists. If the observation of very small earthquakes with $M_w \ll -1$ is supported, then, either the extrapolation of laboratory parameters from metre-scale samples to kilometre-scale faults is problematic, or earthquake initiation involves other processes than simple frictional instability as formulated here. Another strategy for studying earthquake initiation is to examine cases where the immediate trigger of a real earthquake is known.

6.2. Triggering

During the past decade, seismologists have discovered that earthquakes commonly trigger other earthquakes both in the near-field and at distances approaching 4000 km (Hill *et al* 2002). As one of the few cases in nature where the immediate cause of an earthquake is apparent, triggering provides a fundamental clue into initiation. Observed cases of triggering are usually separated into near-field (<2–3 fault lengths) and far-field, with a different set of mechanisms operating in each regime. This distinction may be artificial, as the far-field mechanism must also operate in the near-field, however, it is useful in order to separate plausible regimes for certain physics. Therefore, we will retain the separation here with the above caveats.

6.2.1. Observations. Large earthquakes are followed by abundant smaller earthquakes called aftershocks (section 6.2.4). Aftershocks are, therefore, the most commonly observed form of earthquake interactions. Aftershocks form a cloud around the mainshock rupture plane that can extend up to two fault lengths away. Beginning in the early 1990s, studies such as those by King *et al* (1994) investigated the proposal that aftershocks are triggered by the static stress changes due to the dislocation of the earthquake. As discussed in section 3.1, the deformation of the crust by a slip on a fault plane generates an elastic strain field surrounding the fault. In some areas the strain is extensional, in others it is compressional. The pattern of dilatational strain can most easily be seen for the simple example of a strike-slip fault (figure 28, right).

In addition to the dilatational strain, there is also a deviatoric stress component to the stress field. It is a combination of the shear and normal stress that will determine if a given fault plane slips. Following the Anderson, Hubbert and Rubey failure criterion laid out in section 2, we define the Coulomb stress change $\Delta\tau_c$ on a fault by

$$\Delta\tau_c = -\mu(\Delta\sigma_n - \Delta p) + \Delta\tau, \tag{6.8}$$

where $\Delta\sigma_n$ and $\Delta\tau$ are the resolved normal and shear stress changes, respectively, on a given fault orientation, Δp is the pore pressure change and μ is the coefficient of friction. If $\Delta\tau_c$ increases, then frictional fault slip is promoted (see equation (2.6)). The King *et al* strategy for studying aftershocks is to map the calculated Coulomb stress change based on the observed slip during an earthquake and compare the resulting field with the observed aftershock distribution (figure 28).

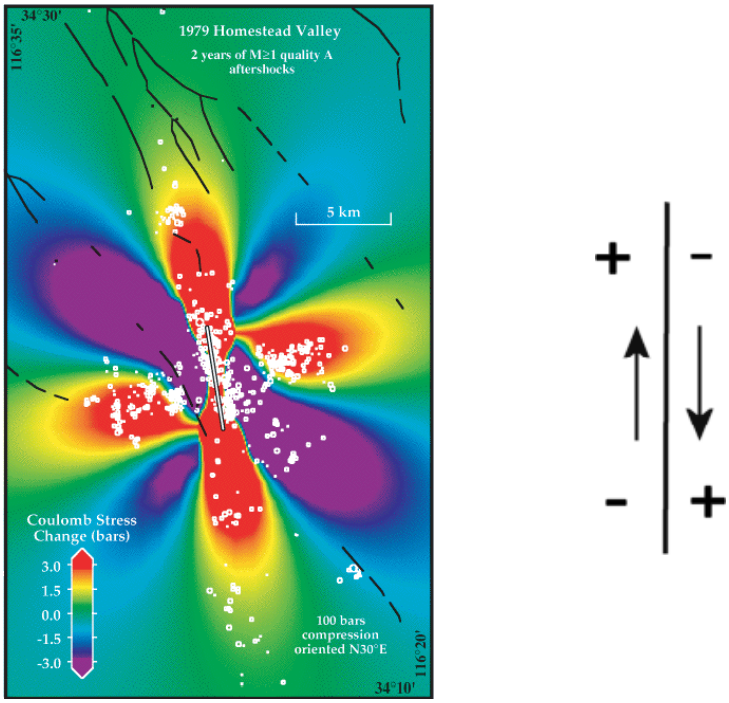

Figure 28. Change in Coulomb stress from the 1979 Homestead Valley mainshock and the subsequent aftershocks. Red (positive) indicates that optimally oriented faults are stressed more towards failure and purple (negative) indicates that failure is inhibited. White circles are observed aftershocks (from King *et al* (1994)). Shown on the right is a schematic of a strike-slip fault with slip in the directions shown by the arrows. The dilatational strain is compressional where there are '+' signs and extensional where there are '−' signs.

This method has had some success in predicting the location of aftershocks and even a few large, nearby earthquakes (Stein 1999). Approximately 85% of the aftershocks of the 1992 $M_w = 7.3$ Landers earthquake occurred where the Coulomb stress field increased at the time of the mainshock (Hardebeck *et al* 1998). A recent review can be found in Harris (2002).

Equation (6.8) by itself does not fully describe the aftershock field shown in figure 28. There are some aftershocks in the areas where failure should have been inhibited by the mainshock. This problem of a continual low aftershock rate in the destressed regions was addressed by adding rate- and state-dependent friction (section 4.2.2) to the stress transfer model (Stein *et al* 2003). Dieterich (1994) showed that if velocity and memory-dependence are incorporated into the standard frictional coefficient based on laboratory experiments, then the rate of seismicity, rather than the absolute number of events, will be influenced by a stress step. Therefore, we might expect some aftershocks to occur in all areas of the stress field if the background seismicity rate is fairly high, but the aftershock rate relative to the background rate will vary systematically with the imposed Coulomb stress.

Since the pattern of static Coulomb stress increase is controlled by the mainshock fault geometry, some other mechanisms could possibly produce similar predictions. In particular, the dynamic stresses follow a very similar pattern because the strong and weak areas of shaking are also determined by the mainshock fault geometry. One major difference between the static and dynamic stress is that only the dynamic field is affected by rupture directivity (section 3.1.5).

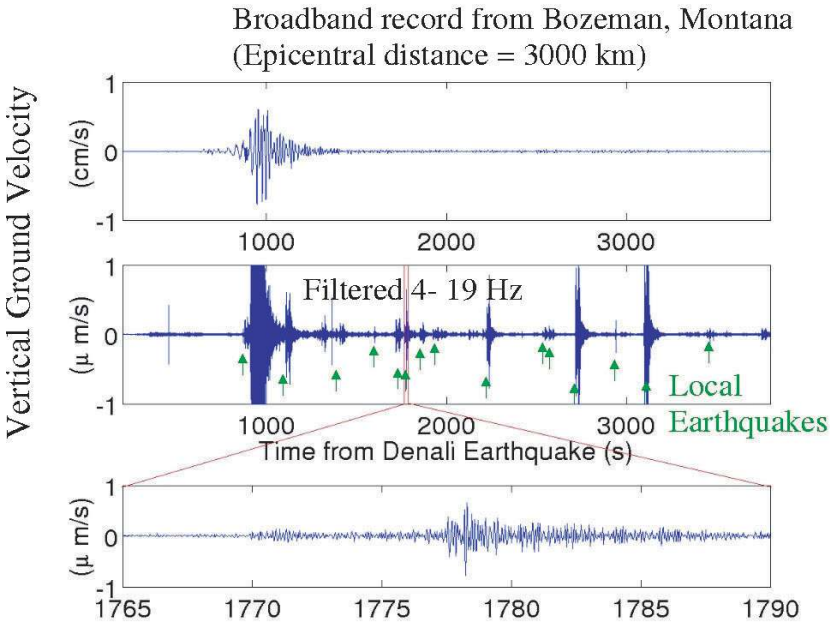

Figure 29. Triggering in Montana from the 2002 Denali, Alaska, earthquake ($M_w = 7.9$). Top panel is the seismic wave generated by the earthquake in Alaska as recorded by a seismometer in Bozeman, Montana. The middle panel is the same record filtered at high-frequencies to show local earthquakes occurring in the vicinity of Montana during the passage of the seismic waves from the Alaskan earthquake. Each green arrow is a local earthquake. The bottom panel is a magnified view of the record of one of these local earthquakes.

Kilb *et al* (2000) and Gomberg *et al* (2003) demonstrated that for earthquakes with strong directivity, the aftershocks are better predicted by the dynamic than the static stress fields. However, the exact mechanism for dynamic triggering is unclear. Furthermore, the dynamic shaking cannot explain the rate decreases, or 'stress shadows', sometimes observed around faults (Stein *et al* 2003). Static stress fields explain stress shadows by invoking negative Coulomb stress changes that move potential faults further from failure. The oscillatory dynamic field has no such negative effect. Areas are simply distinguished by stronger or weaker shaking. Therefore, the current balance of evidence favours static stress as a primary mechanism for generating aftershocks, but the debate is far from over.

Aftershocks can extend up to about 1–2 fault lengths from the original event. Past this distance earthquakes were thought to have no effect until a very surprising observation in 1992. The magnitude 7.3 Landers earthquake in Southern California was followed by up to 10-fold increased seismicity in geothermal and volcanic areas up to 1500 km away, for days after the mainshock (Hill *et al* 1993). Since that time, remote triggered seismicity has been robustly documented for several large events and has now been seen up to 4000 km from the mainshock (Brodsky *et al* 2000, Gomberg *et al* 2001, Eberhardt-Phillips 2003, Prejean *et al* 2004). Figure 29 shows a recent example. In all cases, the triggered sites are geothermal areas. Increases in seismicity have been often observed within the surface wave trains indicating that the seismic waves are the trigger. The triggered seismicity often persists for several days indicating that the seismic waves have a sustained effect on the stress field. The accompanying deformation (Johnston *et al* 1995) also suggests a sustained stress in the triggered regions.

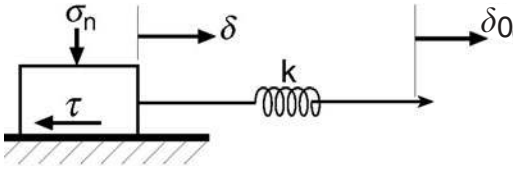

Figure 30. Slider-block sliding on a surface with rate- and state-dependent friction.

Artificially induced seismicity also gives a clue to the initiation process. In addition to the well-controlled Rangely experiment discussed in section 2, humans have produced earthquakes under less advantageous circumstances due to mining, reservoir filling and oil exploitation. Many of these cases are consistent with pore pressure changes relieving the normal stress as was observed for Rangely (Guha 2000). In mines, the excavation also directly removes the load on faults producing the same effect (a recent review is found in McGarr *et al* (2002)).

To summarize, the observations of triggered and induced earthquakes imply that: (1) static stress changes may be effective in the nearfield triggering of aftershocks, (2) seismic waves can trigger earthquakes at long-distances in geothermal areas and (3) pore pressure changes can trigger seismicity. We now explore in detail some theoretical mechanisms for triggering earthquakes that satisfy at least parts of these constraints. At present, no unified earthquake model exists that satisfies all of them.

6.2.2. Triggering with the rate- and state-dependent friction mechanism. If friction on a sliding surface is controlled by the rate- and state-dependent friction law discussed in section 4.2.2, then a sudden change in loading causes a sudden increase in the sliding speed which in turn results in accelerated seismic slip. This mechanism can be important in seismic triggering, as shown by Dieterich (1994). Because this model is widely used in seismology, we discuss this particular mechanism in greater detail than some others.

Consider a slider-block model shown in figure 30 in which friction is controlled by the rate- and state-dependent friction given by (equation (4.24))

$$\mu = \mu_0' + A \ln \dot{\delta} + B \ln \theta. \tag{6.9}$$

Consider the case where $|\theta\dot{\delta}/D_c| \gg 1$ (e.g. large $\dot{\delta}$ during coseismic slip). In this case, from (4.25)

$$\dot{\theta} = -\frac{\theta\dot{\delta}}{D_c}, \tag{6.10}$$

from which we obtain

$$\theta = \theta_0 \exp\left(-\frac{\delta}{D_c}\right) \tag{6.11}$$

and (6.9) becomes

$$\mu = \mu_0' + A \ln \dot{\delta} + B \ln \theta_0 - \frac{B}{D_c}\delta. \tag{6.12}$$

Then, (6.1) and (6.2) become

$$\sigma_n \left(\mu_0' + A \ln \dot{\delta} + B \ln \theta_0 - \frac{B}{D_c}\delta\right) = -k\delta + k\delta_0. \tag{6.13}$$

Spontaneous behaviour. First, we examine the behaviour of this system under constant loading, i.e. $k\delta_0 = \tau_0$. Then, integrating (6.13) with the initial conditions, $\delta = 0$ and $\dot{\delta} = \dot{\delta}_0$ at $t = 0$, we obtain

$$\delta = -\frac{A}{H} \ln \left(1 - \frac{\dot{\delta}_0 H t}{A} \right) \tag{6.14}$$

and

$$\dot{\delta} = \left[\frac{1}{\dot{\delta}_0} - \frac{Ht}{A} \right]^{-1}, \tag{6.15}$$

where

$$H = -\frac{k}{\sigma_n} + \frac{B}{D_c}. \tag{6.16}$$

For an unstable system, $H > 0$. Equations (6.15) shows that the sliding velocity spontaneously increases with time, and at the time

$$t_f = \frac{A}{H} \left(\frac{1}{\dot{\delta}_0} \right) \tag{6.17}$$

$\dot{\delta}$ becomes infinitely large; that is, an instability occurs, i.e. an earthquake occurs. The time t_f is called the time-to-failure.

Loading at a uniform rate. Next, we add loading given by a linear function of time, i.e.

$$k\delta_0 = \tau(t) = \tau_0 + \dot{\tau} t, \tag{6.18}$$

where $\dot{\tau}$ is a constant loading rate. Then, from (6.13)

$$\frac{\tau(t) - k\delta}{\sigma_n} = \left(\mu_0' + A \ln \dot{\delta} + B \ln \theta_0 - \frac{B}{D_c} \delta \right). \tag{6.19}$$

We can integrate (6.19) to obtain,

$$\delta = -\frac{A}{H} \ln \left\{ \frac{\dot{\delta}_0 H \sigma_n}{\dot{\tau}} \left[1 - \exp \left(\frac{\dot{\tau} t}{A\sigma_n} \right) \right] + 1 \right\} \tag{6.20}$$

and

$$\dot{\delta} = \left\{ \left[\frac{1}{\dot{\delta}_0} + \frac{H\sigma_n}{\dot{\tau}} \right] \exp \left(-\frac{\dot{\tau} t}{A\sigma_n} \right) - \frac{H\sigma_n}{\dot{\tau}} \right\}^{-1}. \tag{6.21}$$

The functional forms of δ and $\dot{\delta}$ given by (6.20) and (6.21), respectively, are a little complicated, but both exhibit a monotonic behaviour in time that increases rapidly.

From equation (6.21), the time-to-failure, t_f, is given by

$$t_f = \frac{A\sigma_n}{\dot{\tau}} \ln \left(\frac{\dot{\tau}}{H\sigma_n \dot{\delta}_0} + 1 \right). \tag{6.22}$$

Stepwise change in loading. Next, let us consider the case where a sudden change in loading occurs from τ_0 to τ_1 by $\Delta\tau$ at $t = t_0$ (i.e. $\tau_1 = \tau_0 + \Delta\tau$). This corresponds to the case when a sudden change in the crustal stress occurs due to a large earthquake. Solving equation (6.13) with $k\delta_0 = \tau_0$ for $t < t_0$, and with $k\delta_0 = \tau_0 + \Delta\tau$ for $t > t_0$, and requiring the continuity of slip at $t = t_0$, i.e.

$$\delta(t = t_0 - \varepsilon) = \delta(t = t_0 + \varepsilon),$$

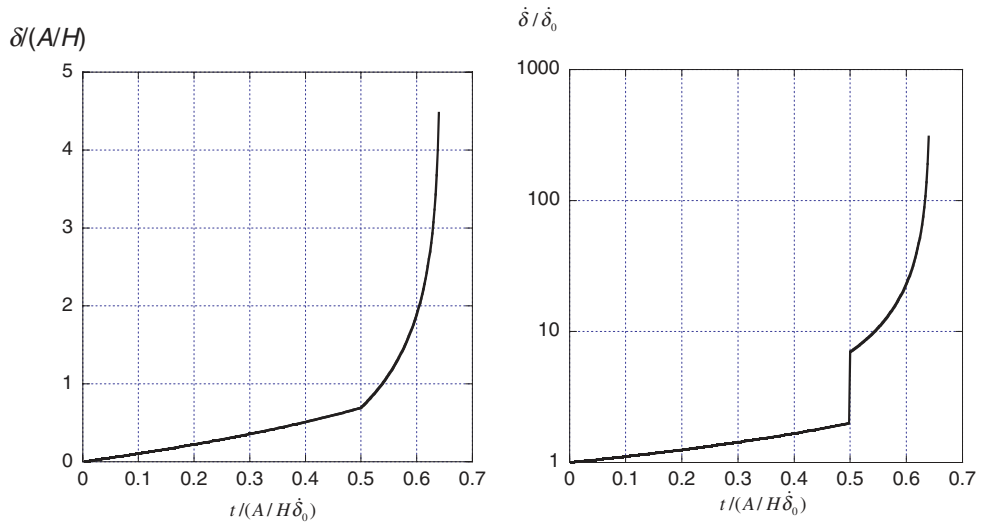

Figure 31. Non-dimensional slip $\delta/(A/H)$ (left) and non-dimensional slip speed $\dot{\delta}/\dot{\delta}_0$ as a function of non-dimensional time, $t/(A/H\dot{\delta}_0)$. A non-dimensional stepwise stress change $\Delta\tau/A\sigma_n = 1.25$ is given at time $t/(A/H\dot{\delta}_0) = 0.5$.

we can derive

$$\dot{\delta}(t = t_0 + \varepsilon) = \dot{\delta}(t = t_0 - \varepsilon) \exp\left(\frac{\Delta\tau}{A\sigma_n}\right). \tag{6.23}$$

Figure 31 shows the non-dimensional slip, $\delta/(A/H)$, and slip speed, $\dot{\delta}/\dot{\delta}_0$, as a function of non-dimensional time $t/(A/H\dot{\delta}_0)$.

Thus, a step-wise change in loading by $\Delta\tau$ causes a step-wise increase in sliding velocity, which in turn causes a step-wise decrease in the time-to-failure. This behaviour, a sudden decrease in the time-to-failure due to a sudden loading, can be used to explain the triggering of seismic activity, and aftershock behaviour (see section 6.2.4).

6.2.3. Triggering with the stress corrosion mechanism. Stress corrosion or sub-critical crack growth is a process widely known in material science (Anderson and Grew 1977, Das and Scholz 1981, Atkinson 1984, Main 1999, Gomberg 2001). Cracks in a purely brittle material remain stable under the loading stress below the critical stress determined by the Griffith criterion. However, under certain environments, especially under high temperatures and with fluids, a crack can grow spontaneously because of weakening near the crack tip due to chemical 'corrosion', even if the loading stress is below the critical level. In this case, a crack is growing constantly, and eventually it will reach a critical state where it fails catastrophically. This mechanism may be important for static triggering of earthquakes in Earth's crust.

Large amounts of experimental data show that the growth rate of a crack with length x is generally given by (Atkinson 1984),

$$\frac{\mathrm{d}x}{\mathrm{d}t} = V_0 \left(\frac{K}{K_0}\right)^p, \tag{6.24}$$

where t is time, K is the stress intensity factor and p is a constant, usually 5 or larger. Although most of the experimental data are obtained for tensile cracks, here we apply this

model to seismic shear cracks. V_0 is the speed of crack growth at $t = 0$, when $K = K_0$. In the following, we assume that $p > 2$. K is given by equation (4.3)

$$K = Yx^{1/2}\sigma, \tag{6.25}$$

where σ is the loading stress, and Y is a constant determined by the geometry of the crack. For a constant loading stress σ, (6.24) can be integrated as

$$x = \frac{x_0}{[1 - ((p-2)/2)(V_0/x_0)t]^{(2/(p-2))}}, \tag{6.26}$$

where x_0 is the crack length at $t = 0$ (Main 1999). This can be rewritten as

$$x = x_0\left(1 + \frac{t}{m\tau}\right)^m, \tag{6.27}$$

where $\tau = (x_0/V_0)$ and $m = 2/(2-p) < 0$. Then,

$$\dot{x} = V_0\left(1 + \frac{t}{m\tau}\right)^{m-1}. \tag{6.28}$$

From equations (6.27) and (6.28), the time-to-failure is given by

$$t_f = -m\tau = -m\frac{x_0}{V_0}. \tag{6.29}$$

Now we consider the case in which the loading stress increases by $\Delta\tau$ at time $t_1 (t_1 < t_f)$. At $t = t_1$, the size and the growth rate of the crack are given by (6.27) and (6.28) as

$$x_1 = x_0\left(1 + \frac{t_1}{m\tau}\right)^m \tag{6.30}$$

and

$$\dot{x}_1 = V_0\left(1 + \frac{t_1}{m\tau}\right)^{m-1} \equiv V_1 \tag{6.31}$$

and at this time, the growth rate suddenly increases with the step-wise increase of the load. From equation (6.24) and (6.25), the relation between the speed just after t_1, V_1^+, and just before t_1, V_1^-, is given by

$$V_1^+ = \left[1 + \frac{\Delta\tau}{\sigma}\right]^p V_1^-. \tag{6.32}$$

Thus, the time-to-failure measured from t_1, is

$$t_f' = -\frac{mx_1}{V_1^+} \tag{6.33}$$

and x and $\dot{x}$ after t_1 are given by

$$x = x_1\left(1 - \frac{t - t_1}{t_f'}\right)^m \tag{6.34}$$

and

$$\dot{x} = V_1^+\left(1 - \frac{t - t_1}{t_f'}\right)^{m-1}. \tag{6.35}$$

As given by equations (6.32) and (6.33), the speed increases suddenly and the time-to-failure is shortened. Figure 32 shows a typical behaviour of x and $\dot{x}$ when the loading is increased stepwise by $\Delta\tau$ at $t = 0.2t_f$.

The stress corrosion behaviour is very similar to that found for the rate- and state-dependent friction law (cf figures 31 and 32.) A sudden increase in loading (e.g. a static stress change due to an earthquake) can accelerate the crack growth, shorten the time to failure, and contribute to increase in the seismicity rate.

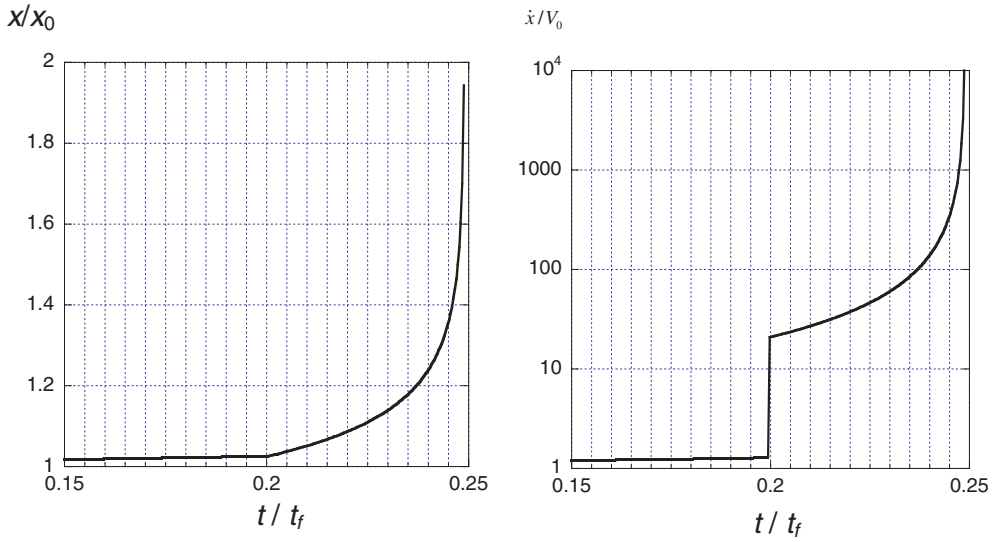

Figure 32. Non-dimensional crack length x/x_0 (left) and non-dimensional crack extension speed $\dot{x}/V_0$ (right) as a function of non-dimensional time $t/t_{\rm f}$. A non-dimensional stepwise stress change $\Delta\tau/\sigma = 0.15$ is imposed at time $t/t_{\rm f} = 0.2$.

6.2.4. Aftershocks and Omori's Law. After a large earthquake (main shock), many smaller earthquakes called aftershocks occur near the rupture zone of the earthquake. As first discovered by Omori (1894), the decay of aftershock activity follows a power law, usually referred to as Omori's Law. (For recent reviews, see Utsu *et al* (1995), Kisslinger (1996), Utsu (2002).)

$$n(t) = \frac{K}{t+c}, \tag{6.36}$$

where $n(t)$ is the number of aftershocks larger than a given magnitude per unit time. A modified (or generalized) Omori's Law is given by

$$n(t) = \frac{K}{(t+c)^p}, \tag{6.37}$$

where p is a constant, which is usually approximately equal to 1. Figure 33 shows two examples. The first one is for the 1891 $M \approx 8$ Nobi, Japan, earthquake, for which Omori found this relationship (Utsu *et al* 1995). It is already more than 100 years (36 500 days) since the mainshock, and we can see that the relation (6.37) holds over a very long period of time ($p = 1$ (constrained), $c = 0.797$ day). The second example is for the 1995 Kobe, Japan, earthquake (Utsu 2002).

Why the aftershock decay follows a power law given by (6.37) has attracted much attention of many seismologists. Many different mechanisms have been proposed, e.g. post seismic creep (e.g. Benioff (1951)), fluid diffusion (Nur and Booker 1972), rate- and state-dependent friction (Dieterich 1994), stress corrosion (Yamashita and Knopoff 1987, Gomberg 2001), etc. Various mechanisms are reviewed in Utsu (1999). In the following, we summarize the two recently developed models, the rate- and state-dependent friction model, and the stress corrosion model.

State- and rate-dependent friction and Omori's Law. Dieterich (1994) assumed that seismicity rate is constant under the constant tectonic loading rate $\dot{\tau}$. This can be accomplished

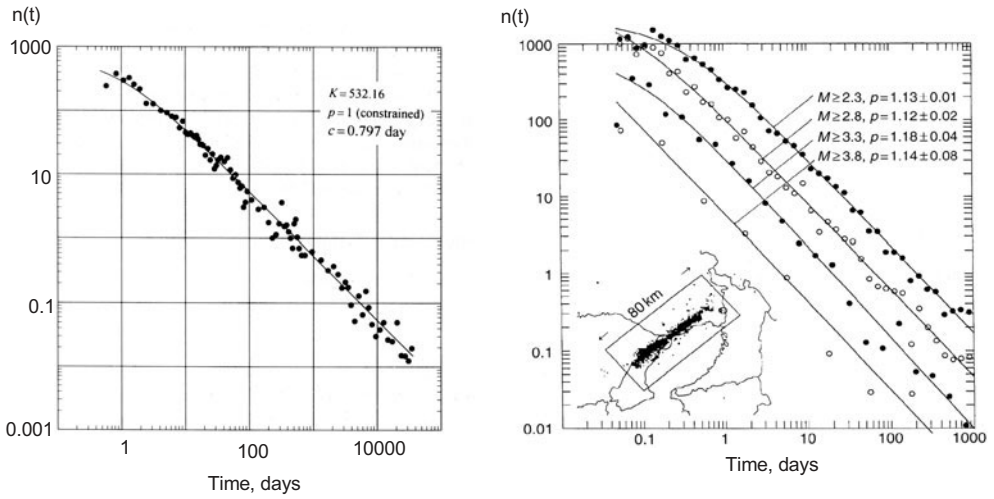

Figure 33. The decay of aftershock activity following the 1891 Nobi, Japan, earthquake, and the 1995 Kobe, Japan earthquake (Utsu 2002).

if earthquake nuclei are distributed such that the time-to-failure, $t_{f0}(n)$, of the nth event is given by $n\Delta t$. In this case, the constant seismicity rate is given by $r_0 = (1/\Delta t)$ (i.e. number of events per unit time). The sliding speed, $\dot{\delta}_{n0}$, of the nth nucleus with $t_{f0}(n)$ can be given by solving equation (6.22) as,

$$\dot{\delta}_{n0} = \frac{\dot{\tau}}{H\sigma} \frac{1}{\exp(\dot{\tau} n \Delta t / A\sigma_n) - 1}. \tag{6.38}$$

If the loading is increased by $\Delta\tau$ due to a mainshock, then, as we discussed in section 6.2.2, the sliding speed increases step-wise by $\exp(\Delta\tau/A\sigma_n)$, the time-to-failure changes, and seismicity rate changes. The new time-to-failure, $t_f(n)$, for nucleus n can be given by substituting the increased sliding velocity into (6.22). Thus,

$$t_f(n) = \frac{A\sigma_n}{\dot{\tau}} \ln\left(\frac{\dot{\tau}}{H\sigma_n \dot{\delta}_{n0} F} + 1\right), \tag{6.39}$$

where $F \equiv \exp(\Delta\tau/A\sigma_n)$. Substituting $\dot{\delta}_{n0}$ given by (6.38) in (6.39), and solving for n, we obtain

$$n = \frac{A\sigma_n}{\dot{\tau}\Delta t} \ln\left\{1 + F\left[\exp\left(\frac{\dot{\tau} t_f(n)}{A\sigma_n}\right) - 1\right]\right\}. \tag{6.40}$$

Here, n and $t_f(n)$ are discrete variables, but we can define the instantaneous seismicity rate R by

$$R = \frac{dn}{dt_f(n)}, \tag{6.41}$$

taking n and $t_f(n)$ as continuous variables.

Thus,

$$\frac{R}{r_0} = \frac{1}{1 - [(F-1)/F]\exp(-\dot{\tau} t / A\sigma_n)}, \tag{6.42}$$

where $r_0 = 1/\Delta t$ is the background rate and $t_f(n)$ is now written as a continuous variable, t. This relation is shown in figure 34.

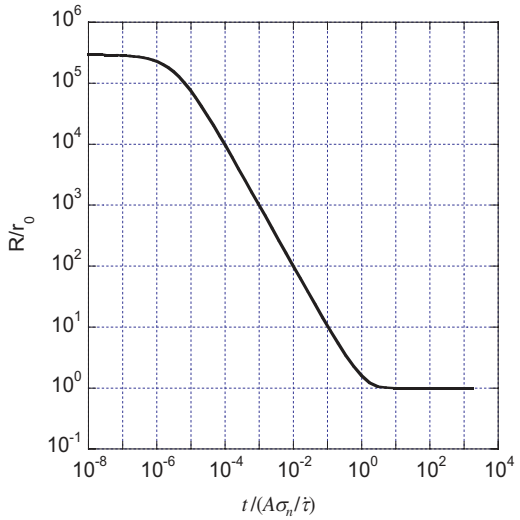

Figure 34. Change in seismicity rate plotted as a function of non-dimensional time $t/(A\sigma_n/\dot{\tau})$, predicted by the rate- and state-dependent friction. The non-dimensional stress change is assumed to be $\Delta\tau/(A\sigma_n) = 12.60$.

For $t \to 0$, $(R/r_0) = F \equiv \exp(\Delta\tau/A\sigma_n)$, which represents the sudden increase in seismicity rate. For times comparable or larger than the timescale of the background stressing $A\sigma_n/\dot{\tau}$, the normalized rate (R/r_0) approaches the steady state value, 1. Between these two extremes, i.e. for $(A\sigma_n/\dot{\tau})/F < t < (A\sigma_n/\dot{\tau})$.

$$\frac{R}{r_0} \approx \frac{a_1}{t + a_2},$$

where $a_1 = (F/(F-1))A\sigma_n/\dot{\tau}$ and $a_2 = (1/(F-1))A\sigma_n/\dot{\tau}$ are constants. This is the form of the Omori's Law.

A test of this model would be whether the observed aftershock decay follows the trends at very small and large t predicted by this model (figure 34). So far, these trends have not been established observationally. The observational difficulties lie in detection thresholds. Immediately after a large earthquake when t is small, many small earthquakes are missed in a catalogue because the larger aftershocks mask their waveforms on seismograms. More progress has been made with the large t limit although it is difficult to measure the duration of a sequence as the detectable measurement is dependent on the choice of spatial windows (Gross and Kisslinger 1997).

Stress corrosion model and Omori's Law. A similar $1/t$ trend can be predicted with the stress corrosion model discussed in section 6.2.3. The results presented here are similar in parts to Gomberg (2001) except that here the equations are formulated to be parallel to the above derivation of Omori's Law from rate- and state-dependent friction.

We assume a constant rate loading in the stress corrosion model,

$$\sigma = \sigma_0 + \dot{\tau}t.$$

Then substituting this in equations (6.24) and (6.25) which, after integration, leads to

$$x = \frac{x_0}{\{1 - (\sigma_0 V_0/2x_0\dot{\tau})((p-2)/(1+p))[(1 + (\dot{\tau}/\sigma_0)t)^{p+1} - 1]\}^{2/(p-2)}} \tag{6.43}$$

and

$$\dot{x} = \frac{V_0(1 + (\dot{\tau}/\sigma_0)t)^p}{\{1 - (\sigma_0 V_0/2x_0\dot{\tau})((p-2)/(1+p))[(1 + (\dot{\tau}/\sigma_0)t)^{p+1} - 1]\}^{2/(p-2)+1}}. \tag{6.44}$$

From (6.44), the time-to-failure t_f can be determined as,

$$t_f = \frac{\sigma_0}{\dot{\tau}} \left\{ \left[\frac{2x_0\dot{\tau}(1+p)}{(p-2)\sigma_0 V_0} + 1 \right]^{1(p+1)} - 1 \right\}. \tag{6.45}$$

As discussed above, a constant seismicity rate $r_0 = (1/\Delta t)$ can be produced by distributing earthquake nuclei such that the time-to-failure, $t_{f0}(n)$, of nth event is given by $n\Delta t$. The sliding speed, V_{n0}, of the nth nucleus with $t_{f0}(n)$ can be given by solving equation (6.45) as,

$$V_{n0} = \frac{2(1+p)}{p-2} \frac{x_0\dot{\tau}}{\sigma_0} \frac{1}{((\dot{\tau}/\sigma_0)n\Delta t + 1)^{p+1} - 1}. \tag{6.46}$$

If the loading is increased by $\Delta\tau$ at $t = 0$ due to a mainshock, then, as we discussed in section 6.2.3 with equation (6.23), the sliding speed increases step-wise by a factor of $F \equiv (1 + (\Delta\tau/\sigma_0))^p$, the time-to-failure changes, and the seismicity rate changes. The new time-to-failure, $t_f(n)$, for a nucleus n can be given by substituting the increased sliding velocity, FV_{n0}, into (6.45). Thus,

$$\begin{aligned} t_f(n) &= \frac{\sigma_0}{\dot{\tau}} \left\{ \left[\frac{2x_0\dot{\tau}(1+p)}{(p-2)\sigma_0 F V_{n0}} + 1 \right]^{1/(p+1)} - 1 \right\} \\ &= \frac{\sigma_0}{\dot{\tau}} \left\{ \left[1 + \frac{((\dot{\tau}/\sigma_0)n\Delta t + 1)^{p+1} - 1}{F} \right]^{1/(p+1)} - 1 \right\}. \end{aligned} \tag{6.47}$$

Solving this for n, we obtain

$$n = \frac{\sigma_0}{\dot{\tau}\Delta t} \left\{ \left[F \left[\left(\frac{\dot{\tau}}{\sigma_0} t_f(n) + 1 \right)^{p+1} - 1 \right] + 1 \right]^{1/(p+1)} - 1 \right\}. \tag{6.48}$$

Here, n and $t_f(n)$ are discrete variables, but we can define the instantaneous seismicity rate R by

$$R = \frac{\mathrm{d}n}{\mathrm{d}t_f(n)} \tag{6.49}$$

taking n and $t_f(n)$ as continuous variables.

Thus,

$$\frac{R}{r_0} = F \left[F + (1-F) \left(1 + \frac{\dot{\tau}}{\sigma_0} t \right)^{-(p+1)} \right]^{-p/(p+1)}, \tag{6.50}$$

where $t_f(n)$ is now denoted simply by t. For $t = 0$, $R/r_0 = F$ and for $t \gg \sigma_0/\dot{\tau}$, $R/r_0 = F^{1/(p+1)}$.

Because p is large, for $t \ll \sigma_0/\dot{\tau}$, R/r_0 has the form $a_1/((a_2 + t)^{p/(1+p)}) \approx a_1/(a_2 + t)$ which is the Omori's law, where $a_1 = (F/(F-1))(\sigma_0/(p+1)\dot{\tau})$ and $a_2 = (1/(F-1))(\sigma_0/(p+1)\dot{\tau})$.

An example of a solution with typical parameters is shown in figure 35. The similarity to the rate and state model solution in figure 34 suggests that the form of the decay in seismicity with time is controlled by the similar mathematical construction with both cases. In both cases, the initial distribution of earthquake nuclei is designed to yield a constant rate under constant stressing rates.

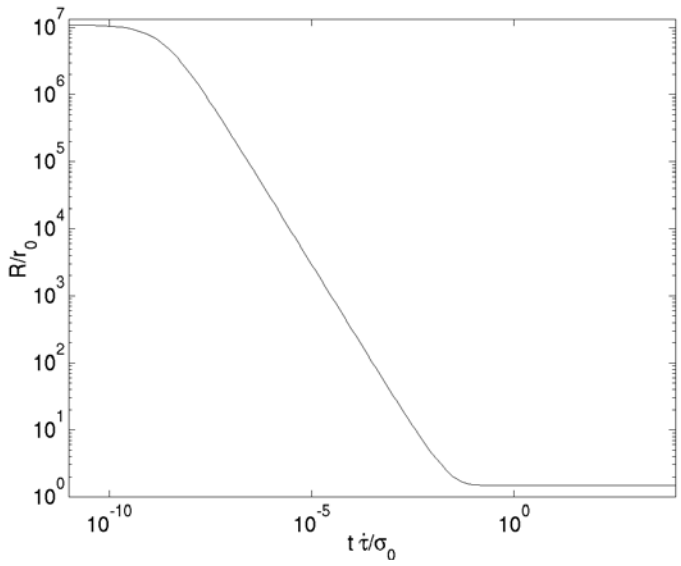

Figure 35. Change in seismicity rate plotted as a function of non-dimensional time $t/(\sigma_0/\dot{\tau})$, predicted by the stress corrosion model. The parameters used are $p = 40$, and $\Delta\tau/\sigma = 0.5$.

One difference between the stress corrosion and the rate- and state-dependent friction model is that, in the former, the rate returns exactly to the background rate only for infinite p while, in the latter, it always returns to the background rate. If aftershocks are generated by sub-critical crack growth, then we must rely on long-term relaxation processes such as viscous flow to prevent a continual ratcheting upwards of seismicity.

6.2.5. Hydrologic barrier removal. The above mechanisms emphasize the solid mechanics of earthquake initiation. As discussed in section 2 and illustrated by the observations of artificially induced seismicity, fluid movement can reduce the strength on faults and initiate earthquakes. Recent research has begun exploring quantitative models and new observational techniques in order to constrain fluid movements and their importance in natural faults. One example of a fluid-based triggering mechanism comes from recent work on the removal of transient hydrologic barriers during ground shaking (Brodsky *et al* 2003). Seismic waves can induce water flow into faults as the differential stiffness of geological units generates a hydraulic gradient when the seismic waves impose a long-wavelength, oscillating strain field. Even very small fluid shear stresses $\sim$1 Pa are sufficient to remove accumulations of sediment or precipitate (Kessler 1993). The sediment or precipitate barriers blocked flow prior to the earthquake while maintaining a sharp pressure differential Δp which, according to the standard formulation of flow in porous media (Darcy's Law), is of the order of

$$\Delta p = \frac{U_{\mathrm{d}} h \eta}{k},$$

where U_{d} is the average fluid flow velocity (Darcy velocity), h is the thickness of the barrier, η is the viscosity of the water and k is the permeability of the rock (e.g. Freeze and Cherry (1979)). When the earthquake occurs, the removal of the barrier redistributes this pressure difference Δp in the fault zone. In places where the pressure rises, the frictional stress is reduced (see equation (2.6)), and failure can occur. For realistic parameters, the pressure changes can be

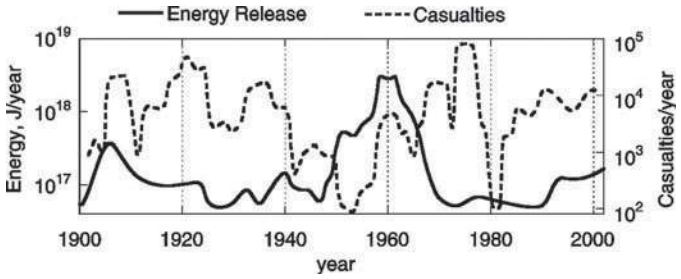

Figure 36. Energy release and casualties (number of death) per year during the 20th century. Five-year running average is taken. The peak in the energy release in 1960 is mainly due to a sequence of great earthquakes from 1952 to 1964 including the 1960 Chilean earthquake. The peak in the casualty in the mid 1970s is due to the 1976 Tanshang, China, earthquake (updated from Kanamori (1978)).

0.04 MPa (Brodsky *et al* 2003), which is sufficient to trigger earthquakes based on static stress studies of triggering thresholds in the nearfield (Hardebeck *et al* 1998).

It is obvious that the fluid-based models are not nearly as well developed quantitatively as the solid models like rate- and state-dependent friction and stress corrosion. Much more theoretical, observational and experimental work is necessary to develop the formalism to compare both the fluid and solid avenues for triggering. However, even at this stage the fluid approach is able to address some problems, such as sustained distant triggering, that elude the solid models. Neither stress corrosion nor rate- and state-dependent friction can explain sustained triggering from seismic waves (Gomberg 2001). We look forward to the development of this promising line of research.

7. Conclusions

We have presented in this paper an overview of earthquake physics with an emphasis on initiation. There are many interesting and active areas of research that we have omitted because of space (for recent reviews, see Lee *et al* (2002, 2003), National Research Council (2003)). Subjects that we neglected include questions of how earthquakes stop, geological studies of earthquake occurrence histories and methods of mitigating earthquake damage. The latter area is particularly important for society. Figure 36 shows that casualties from earthquakes are not correlated to magnitude or overall level of seismicity. They are more closely related to the engineering infrastructures and preparedness of a region.

The above discussion should have made it clear that the quantitative prediction of earthquake initiation is an extremely complicated and perhaps impossible task. Even in the best case scenario of a predictable fault nucleation length, the nucleation length of 1 m requires instruments to be too densely spaced to be practical. Perhaps one day we will be able to accomplish accurate earthquake prediction, but the current state of the science implies that that day is decades, if not centuries away. In the short term, it is more practical to save lives by using the detailed knowledge we have about the propagation of seismic waves and strength of seismic shaking to design buildings and infrastructure that will protect people during an earthquake. Recent engineering advances such as active- and passive-controlled buildings and dense, high-quality ground motion monitoring brings the goal of saving lives well within our grasp.

At the same time, we continue to build a basic scientific framework to learn why and how earthquakes begin. Over the last 10–20 years we have unravelled parts of the puzzle based on the state of stress in the crust, detailed slip inversions, laboratory friction models,

complex system modelling and triggering studies. A number of still unanswered questions remain. Many of the questions highlighted during the course of this review will only be addressed by improved instrumentation and observational techniques. Technical developments like the recently deployed dense seismic networks of Japan and the Earthscope instrumentation initiative in the United States may help us measure seismic parameters like stress drop and radiated energy to sufficient accuracy to finally address the dynamic similarity of earthquakes. Another recent advance is the use of non-seismic data to study earthquakes. This review emphasized elastic waves as the most developed method for characterizing earthquakes, but insights from geodetic methods are becoming increasingly important. As the subject expands beyond classic elasticity, we predict that input from geology and hydrogeology will play a greater role in the decades to come. Combining disciplines might allow us to measure migrating fluids and if their pressures can equal or exceed the minimum principal stress as suggested by normal faulting studies. We must also grapple with the thickness and physical properties of the fault zone. Seismic and other observations will also have to address the heterogeneity of stress and strength in the crust. Mechanisms for triggering can be differentiated by studies targeting phenomena where the predictions diverge. The beginning of aftershock sequences, the existence of very small earthquakes and the occurrence of long-range triggering are a few areas with some resolving power.

Theorists and experimentalists also have their work cut out for them. Averages like that taken in estimating the stress drop need to put on firmer theoretical ground. Theorists and experimentalists will have to explore the relationship between initiation conditions and rupture propagation. Is the same physics applicable or does a new set of processes come into play once rapid slip has begun? What is the physical nature of the fracture energy term that controls dynamics ranging from slow quasi-static slip to brittle failure with high rupture speed and efficient energy radiation? What happens when the mechanisms such as thermal pressurization and lubrication are combined in a single rupture model? The complexity community must try to ascertain exactly how chaotic are earthquakes. Even chaotic systems, like weather, can be predicted over short time horizons if the observations have sufficient resolution. Quantification of the required resolution and the divergence rate from initial conditions would be a valuable contribution.

As our observational database improves, our computational ability accelerates and our laboratories become more refined, the next few decades promise to bring more earthquake insights and perhaps some answers.

Acknowledgments

We thank Allan Rubin and an anonymous reviewer for constructive comments.

References

Abercrombie R 1995 *J. Geophys. Res.* **100** 24003–14
Abercrombie R and Leary P 1993 *Geophys. Res. Lett.* **20** 1511–4
Aki K 1966 *Bull. Earthquake Res. Inst. Tokyo Univ.* **44** 73–88
Aki K and Richards P G 2002 *Quantitative Seismology* (Sausalito: University Science Books) p 685
Anderson E M 1905 *Trans. Edin. Geol. Soc.* **8** 393
Anderson E M 1951 *The Dynamics of Faulting and Dyke Formation with Applications to Britain* (Edinburgh: Oliver and Boyd) p 206
Anderson O L and Grew P C 1977 *Rev. Geophys.* **15** 77–104
Andrews D J 2002 *J. Geophys. Res.* **107** (B2) 2363
Atkinson B K 1984 *J. Geophys. Res.* **89** 4077–114

Axen G J 2004 Low-angle normal fault mechanics and crustal strength *Rheology and Deformation in the Lithosphere at Continental Margins* ed G Karner *et al* (New York: Columbia University Press) pp 46–91
Bak P and Tang C 1989 *J. Geophys. Res.* **94** 15635–7
Bak P *et al* 1988 *Phys. Rev.* A **38** 364–74
Barenblatt G I 1962 The mathematical theory of equilibrium cracks in brittle fracture *Advances in Applied Mechanics* vol 7 (London: Academic) pp 55–125
Beeler N M *et al* 1994 *Geophys. Res. Lett.* **21** 1987–90
Benioff H 1951 *Bull. Seismol. Soc. Am.* **41** 31–62
Beroza G C and Mikumo T 1996 *J. Geophys. Res.* **101** 22449–60
Bird P and Kagan Y Y 2004 *Bull. Seismol. Soc. Am.* at press
Boatwright J and Choy G L 1986 *J. Geophys. Res.* **91** 2095–112
Boatwright J *et al* 2002 *Bull. Seismol. Soc. Am.* **92** 1241–55
Bouchon M 1997 *J. Geophys. Res.* **102** 11731–44
Bouchon M and Vallee M 2003 *Science* **301** 824–6
Brace W F and Byerlee J D 1966 *Science* **153** 990–2
Brodsky E E and Kanamori H 2001 *J. Geophys. Res.* **106** 16357–74
Brodsky E E *et al* 2000 *Geophys. Res. Lett.* **27** 2741–4
Brodsky E E *et al* 2003 *J. Geophys. Res.* **108**
Brune J N 1970 *J. Geophys. Res.* **75** 4997–5009
Burridge R and Knopoff L 1964 *Bull. Seismol. Soc. Am.* **54** 1875–88
Burridge R and Knopoff L 1967 *Bull. Seismol. Soc. Am.* **57** 341–71
Byerlee J 1978 *Pure Appl. Geophys.* **116** 615–26
Byerlee J 1992 *Tectonophysics* **211** 295–303
Cardwell R K *et al* 1978 *Geophys. J. R. Astron. Soc.* **52** 525–30
Chester F M and Chester J S 1998 *Tectonophysics* **295** 199–221
Choy G L and Boatwright J L 1995 *J. Geophys. Res.* **100** 18205–28
Dahlen F A 1977 *Geophys. J. R. Astron. Soc.* **48** 239–61
Dahlen F A and Tromp J 1998 *Theoretical Global Seismology* (Princeton, NJ: Princeton University Press) p 1025
Das S 1988 *Bull. Seismol. Soc. Am.* **78** 924–30
Das S and Scholz C H 1981 *J. Geophys. Res.* **86** 6039–51
Dieterich J 1994 *J. Geophys. Res.* **99** 2601–18
Dieterich J H 1979 *J. Geophys. Res.* **84** 2161–8
Dmowska R and Rice J R 1986 Fracture theory and its seismological applications *Theories in Solid Earth Physics* ed R Teisseyre (Warzawa: PWN-Polish Publishers) pp 187–255
Dragert H *et al* 2001 *Science* **292** 1525–8
Dugdale D S 1960 *J. Mech. Phys. Solids* **8** 100–4
Eberhart-Phillips D *et al* 2003 *Science* **300** 1113–18
Ekström E 2000 Global studies of earthquakes *Problems in Geophysics for the New Millennium* ed G E E Boschi and A Morelli (Bologna: Editorice Compositori) pp 111–24
Eshelby J D 1969 *J. Mech. Phys. Solids* **17** 177–99
Fossum A F and Freund L B 1975 *J. Geophys. Res.* **80** 3343–7
Freeze R A and Cherry J A 1979 *Groundwater* (Englewood Cliffs, NJ: Prentice Hall)
Freund L B 1972 *J. Elasticity* **2** 341–9
Freund L B 1989 *Dynamic Fracture Mechanics* (Cambridge: Cambridge University Press) p 563
Gilbert F and Dziewonski A M 1975 *Phil. Trans. R. Soc. Lond.* **278** 187–269
Gomberg J 2001 *J. Geophys. Res.* **106** 16253–63
Gomberg J *et al* 2001 *Nature* **411** 462–6
Gomberg J *et al* 2003 *Bull. Seismol. Soc. Am.* **93** 118–38
Griffith A A 1920 *Phil. Trans. R. Soc. Lond.* A **221** 169–98
Gross S and Kisslinger C 1997 *J. Geophys. Res. Solid Earth* **102** 7603–12
Guha S K 2000 *Induced Earthquakes* (London: Kluwer) p 314
Gutenberg B and Richter C F 1941 *Geol. Soc. Am.* **34** 1–131 (special paper)
Hardebeck J L and Hauksson E 2001 *J. Geophys. Res.* **106** 21859–82
Hardebeck J L *et al* 1998 *J. Geophys. Res.* **103** 24427–37
Harris R A 2002 Stress triggers, stress shadows, and seismic hazard *International Handbook of Earthquake & Engineering Seismology* part B, ed H Kanamori *et al* (San Diego, CA: Academic) pp 1217–32
Haskell N 1964 *Bull. Seismol. Soc. Am.* **56** 1811–42
Heaton T 1990 *Phys. Earth Planet. Inter.* **64** 1–20

Hergaten S 2002 *Self-Organized Criticality in Earth Systems* (Berlin: Springer) p 272
Hill D P *et al* 2002 *Phys. Today* **55** 41–7
Hill D P *et al* 1993 *Science* **260** 1617–23
Houston H 2001 *J. Geophys. Res.* **106** 11137–50
Hubbert M K and Rubey W W 1959 *Geol. Soc. Am.* **70** 115–66 (Special paper)
Husseini M I 1977 *Geophys. J. R. Astron. Soc.* **49** 699–714
Ida Y 1972 *J. Geophys. Res.* **77** 3796–805
Ide S and Beroza G C 2001 *Geophys. Res. Lett.* **28** 3349–52
Ide S and Takeo M 1997 *J. Geophys. Res.* **102** 27379–91
Ikeda R *et al* 2001 *Isl. Arc.* **10** 252–60
Izutani Y and Kanamori H 2001 *Geophys. Res. Lett.* **28** 4007–10
Jaeger J C and Cook N G W 1979 *Fundamentals of Rock Mechanics* (London: Chapman and Hall) p 593
Jeffreys H 1942 *Geol. Mag.* **79** 291–5
Johnston M J S *et al* 1995 *Bull. Seismol. Soc. Am.* **85** 787–95
Jost M L *et al* 1998 *Bull. Seismol. Soc. Am.* **88** 815–32
Kadanoff L P 1991 *Phys. Today* **44** 9–10
Kanamori H 1978 *Nature* **271** 411–14
Kanamori H 1983 Global seismicity *Earthquakes: Observation, Theory and Interpretation* ed H Kanamori and E Boschi (New York: North-Holland) pp 596–608
Kanamori H and Anderson D L 1975 *Bull. Seismol. Soc. Am.* **65** 1073–95
Kanamori H *et al* 1998 *Science* **279** 839–42
Kanamori H and Brodsky E E 2001 *Phys. Today* **54** 34–40
Kanamori H *et al* 1993 *Bull. Seismol. Soc. Am.* **83** 330–46
Kessler J H 1993 *Berkeley* University of California, Berkeley
Kikuchi M 1992 *Tectonophysics* **211** 107–13
Kikuchi M and Fukao Y 1988 *Bull. Seismol. Soc. Am.* **78** 1707–24
Kikuchi M and Kanamori H 1994 *Geophys. Res. Lett.* **21** 2341–4
Kikuchi M and Kanamori H 1995 *Pure Appl. Geophys.* **144** 441–53
Kilb D *et al* 2000 *Nature* **408** 570–4
King G C P *et al* 1994 *Bull. Seismol. Soc. Am.* **84** 935–53
Kinoshita S and Ohike M 2002 *Bull. Seismol. Soc. Am.* **92** 611–24
Kisslinger C 1996 Aftershocks and fault-zone properties *Advances in Geophysics* vol 38, ed R Dmowska and B Saltzman (San Diego, CA: Academic) pp 1–36
Knopoff L 1958 *Geophys. J.* **1** 44–52
Kostrov B V 1966 *J. Appl. Math. Mech.* **30** 1241–8 (PMM) (Engl. tans.)
Kostrov B V 1974 *Izv. Earth Phys.* **1** 23–40
Lachenbruch A H 1980 *J. Geophys. Res.* **85** 6097–112
Lachenbruch A H and Sass J H 1980 *J. Geophys. Res.* **85** 6185–222
Lapusta N and Rice J R 2003 *J. Geophys. Res.* **108** (B4) 2205
Lawn B 1993 *Fracture of Brittle Solids* 2nd edn (Cambridge: Cambridge University Press) p 378
Lay T and Wallace T C 1995 *Modern Global Seismology* (San Diego, CA: Academic) p 521
Lee W H K *et al* 2002 *International Handbook of Earthquake & Engineering Seismology* part A (San Diego, CA: Academic) pp 1–933
Lee W H K *et al* 2003 *International Handbook of Earthquake & Engineering Seismology* part B (San Diego, CA: Academic) pp 937–1945
Li V C 1987 Mechanics of shear rupture applied to earthquake zones *Fracture Mechanics of Rock* (London: Acdemic) pp 351–428
Linker M and Dieterich J H 1992 *J. Geophys. Res. Solid Earth* **97** 4923–40
Lockner D A 1995 Rock failure in rock physics & phase relations *A Handbook of Physical Constants* vol 3, ed T J Ahrens (Washington, DC: American Geophysical Union) pp 127–47
Lockner D A and Beeler N M 2002 Rock failure and earthquakes *International Handbook of Earthquake & Engineering Seismology* part A, ed W H K Lee *et al* (San Diego, CA: Academic) pp 505–37
Ma K F *et al* 2003 *Geophys. Res. Lett.* **30** (5) 1244
Ma K-F *et al* 1999 *EOS Trans.—AGU* **80** 605
Madariaga R 1977 *Pure Appl. Geophys.* **115** 301–16
Madariaga R 1979 *J. Geophys. Res.* **84** 2243–50
Madariaga R and Olsen K B 2002 Earthquake dynamics *International Handbook of Earthquake & Engineering Seismology* part A, ed W H K Lee *et al* (San Diego, CA: Academic) pp 175–94

Main I 1996 *Rev. Geophys.* **34** pp 433–62
Main I G 1999 *Geophys. J. Int.* **139** F1–6
Marder M and Fineberg J 1996 *Phys. Today* **49** 24–9
Marone C and Kilgore B 1993 *Nature* **362** 618–21
Maruyama T 1964 *Bull. Earthquake Res. Inst. Tokyo Univ.* **42** 289–368
Mase C W and Smith L 1985 *Pure Appl. Geophys.* **122** 583–607
Mase C W and Smith L 1987 *J. Geophys. Res.* **92** 6249–72
Massonnet D *et al* 1993 *Nature* **364** 138–42
Mayeda K and Walter W R 1996 *J. Geophys. Res.* **101** 11195–208
McGarr A and Fletcher J B 2002 *Bull. Seismol. Soc. Am.* **92** 1633–46
McGarr A *et al* 2002 Case histories of induced and triggered seismicity *International Handbook of Earthquake & Engineering Seismology* part A, ed W H K Lee *et al* (San Diego, CA: Academic) pp 647–61
McKenzie D P and Brune J N 1972 *Geophys. J. R. Astron. Soc.* **29** 65–78
Mikumo T and Miyatake T 1993 *Geophys. J. Int.* **112** 481–96
Mikumo T *et al* 2003 *Bull. Seismol. Soc. Am.* **93** 264–82
Miyatake T 1992 *Geophys. Res. Lett.* **19** 1041–4
Mott N F 1948 *Engineering* **165** 16
National Research Council 2003 *Living on an Active Earth* (Washington, DC: National Academies) p 418
Nur A and Booker J 1972 *Science* **175** 885–7
Ohnaka M and Shen L-F 1999 *J. Geophys. Res.* **104** 817–44
Omori F 1894 *J. College of Science Imperial University of Tokyo* **7** 111–200
Otsuka M 1971 *J. Seismol. Soc. Japan* **24** 215–27
Palmer A C and Rice J R 1973 *Proc. R. Soc. Lond. Ser. A—Math. Phys. Eng. Sci.* **332** 527–48
Perez-Campos X and Beroza G C 2001 *J. Geophys. Res.* **106** 11127–36
Prejean S G and Ellsworth W L 2001 *Bull. Seismol. Soc. Am.* **91** 165–77
Prejean S G *et al* 2004 *Bull. Seismol. Soc. Am.* **94** at press
Provost A S and Houston H 2003 *J. Geophys. Res.* **108** (B3) 2175
Quin H 1990 *Tectonophysics* **175** 93–117
Rabinowicz E 1995 *Friction and Wear of Materials* (New York: Wiley) p 315
Raleigh C B *et al* 1976 *Science* **191** 1230–7
Rice J R 1980 The mechanics of earthquake rupture *Physics of the Earth's Interior* ed A M Dziewonski and E Boschi (Amsterdam: North-Holland) pp 555–649
Rice J R 1992 Fault stress states, pore pressure distributions, and the weakness of the San Andreas fault *Fault Mechanics and Transport Properties of Rock* ed B Evans and T-F Wong (San Diego, CA: Academic) pp 475–503
Richards P G 1976 *Bull. Seismol. Soc. Am.* **66** 1–32
Richardson E and Jordan T H 2002 *Bull. Seismol. Soc. Am.* **92** 1766–82
Romanowicz B and Ruff L J 2002 *Geophys. Res. Lett.* **29** (12) 1604
Romanowicz B and Rundle J B 1993 *Bull. Seismol. Soc. Am.* **83** 1294–7
Rosakis A J 2002 *Adv. Phys.* **51** 1189–257
Rubin A M *et al* 1999 *Nature* **400** 635–41
Rudnicki J W and Kanamori H 1981 *J. Geophys. Res.* **86** 1785–93
Ruina A L 1983 *J. Geophys. Res.* **88** 10359–70
Rundle J B *et al* 2000 *GeoComplexity and the Physics of Earth* (Washington, DC: American Geophysical Union) p 284
Saffer D M *et al* 2003 *J. Geophys. Res.* **108** (B5) 2274
Sass J H *et al* 1997 *J. Geophys. Res.* **102** 27575–85
Scholz C H 1994 *Bull. Seismol. Soc. Am.* **84** 215–18
Scholz C H 2002 *The Mechanics of Earthquakes and Faulting* (New York: Cambridge University Press) p 471
Sibson R H 1973 *Nature* **243** 66–8
Sibson R H 1985 *J. Struct. Geol.* **7** 751–4
Sibson R H 2003 *Bull. Seismol. Soc. Am.* **93** 1169–78
Sibson R H and Xie G Y 1998 *Bull. Seismol. Soc. Am.* **88** 1014–22
Simons M *et al* 2002 *Bull. Seismol. Soc. Am.* **92** 1390–402
Singh S K *et al* 2004 *Bull. Seismol. Soc. Am.* **94** at press
Stein R S 1999 *Nature* **402** 605–9
Stein R S *et al* 2003 *Sci. Am.* **288** 72–9
Stekette J A 1958 *Can. J. Phys.* **36** 1168–98
Tibi R *et al* 2003 *J. Geophys. Res.* **108** (B2) 2091

Tsuboi C 1933 *Japan. J. Astron. Geophys.* **10** 93–248
Turcotte D L 1997 *Fractals and Chaos in Geology and Geophysics* (Cambridge, UK: Cambridge University Press) p 398
Turcotte D L and Malamud B D 2002 Earthquakes as a complex system *International Handbook of Earthquake & Engineering Seismology* part A, ed W H K Lee *et al* (San Diego, CA: Academic) pp 209–35
Utsu T 1999 *Seismicity Studies: A Comprehensive Review* (Tokyo: University of Tokyo Press) p 876 (in Japanese)
Utsu T 2002 Statistical features of seismicity *International Handbook of Earthquake & Engineering Seismology* part A, ed W H K Lee (San Diego, CA: Academic) pp 719–32
Utsu T *et al* 1995 *J. Phys. Earth* **43** 1–33
Uyeda S 1978 *The New View of the Earth* (San Francisco: W H Freeman and Company) p 217
Venkataraman A and Kanamori H 2004 *J. Geophys. Res.* **109** B05302, 10.1029/2003JB002549
Venkataraman A *et al* 2002 *Bull. Seismol. Soc. Am.* **92** 1256–65
Wald D J and Heaton T H 1994 *Bull. Seismol. Soc. Am.* **84** 668–91
Wernicke B 1981 *Nature* **291** 645–8
Wyss M and Brune J N 1968 *J. Geophys. Res.* **73** 4681–94
Yamashita T and Knopoff L 1987 *Geophys. J.* **91** 13–26
Zeng Y H *et al* 1994 *Geophys. Res. Lett.* **21** 725–8
Zoback M D and Healy J H 1984 *Ann. Geophys.* **2** 689–98
Zoback M D *et al* 1987 *Science* **238** 1105–11