From Geodetic Imaging of Seismic and Aseismic Fault Slip to Dynamic Modeling of the Seismic Cycle

Jean-Philippe Avouac

Bullard Laboratories, Department of Earth Sciences, University of Cambridge, Cambridge CB3 0EZ, United Kingdom; email: jpha2@cam.ac.uk

Keywords
earthquake, aseismic creep, seismic and interseismic coupling, asperities

Abstract
Understanding the partitioning of seismic and aseismic fault slip is central to seismotectonics as it ultimately determines the seismic potential of faults. Thanks to advances in tectonic geodesy, it is now possible to develop kinematic models of the spatiotemporal evolution of slip over the seismic cycle and to determine the budget of seismic and aseismic slip. Studies of subduction zones and continental faults have shown that aseismic creep is common and sometimes prevalent within the seismogenic depth range. Interseismic coupling is generally observed to be spatially heterogeneous, defining locked patches of stress accumulation, to be released in future earthquakes or aseismic transients, surrounded by creeping areas. Clay-rich tectonites, high temperature, and elevated pore-fluid pressure seem to be key factors promoting aseismic creep. The generally logarithmic time evolution of afterslip is a distinctive feature of creeping faults that suggests a logarithmic dependency of fault friction on slip rate, as observed in laboratory friction experiments. Most faults can be considered to be paved with interlaced patches where the friction law is either rate-strengthening, inhibiting seismic rupture propagation, or rate-weakening, allowing for earthquake nucleation. The rate-weakening patches act as asperities on which stress builds up in the interseismic period; they might rupture collectively in a variety of ways. The pattern of interseismic coupling can help constrain the return period of the maximum-magnitude earthquake based on the requirement that seismic and aseismic slip sum to match long-term slip. Dynamic models of the seismic cycle based on this conceptual model can be tuned to reproduce geodetic and seismological observations. The promise and pitfalls of using such models to assess seismic hazard are discussed.
1. INTRODUCTION

How much, where, when, and why does aseismic creep take place on a fault? Answering these questions is essential both in quantitatively assessing the seismic potential of a given fault and in laying the foundation of a physics-based theory of earthquake behavior that could be used in practice for seismic hazard assessment. The advent of space-based geodesy and remote sensing techniques to monitor crustal deformation has provided new tools to investigate these issues. In this article, I discuss the insights gained from combining these techniques with seismological and geological methods. We will see that, in favorable cases, it is possible to document the partitioning between seismic and aseismic slip in time and space. I review such examples and discuss how these observations can be interpreted in view of dynamic models of fault slip.

2. PRINCIPLES OF FAULT SLIP BUDGET DETERMINATION

2.1. Relating Seismic and Aseismic Slip with Surface Deformation

I start by describing a simple two-dimensional kinematic model of the seismic cycle for the particular case of a subduction megathrust (a thrust fault of crustal to lithospheric scale). This simple kinematic model is a useful reference to which observations and numerical simulations can be compared (Figure 1). I examine here only the Sumatra example of a subduction megathrust, but the model can easily be adapted to other tectonic contexts. This model conforms to early views of the seismic cycle along subduction (e.g., Savage 1983) or continental (e.g., Cattin & Avouac 2000) megathrusts by assuming that the fault is fully locked over the seismogenic depth range. The term cycle does not imply a periodic behavior but refers instead to the process by which earthquakes recurrently nucleate, grow, and arrest, releasing elastic strain accumulated over intervening interseismic periods.

A fault can be locked, or it can slip seismically, radiating seismic waves, or aseismically. Aseismic slip can be relatively steady (interseismic creep), or it can be transient, resulting from postseismic relaxation (afterslip) or spontaneous slip events (termed slow slip events, or SSEs) (Schwartz & Rokosky 2007). At every point on the interface, seismic and aseismic slip must sum in the long term to the displacement required by the long-term average slip rate. Slip can result from a recurring characteristic earthquake (Schwartz & Coppersmith 1984) with associated afterslip, from earthquakes with variable magnitude and associated afterslip, or from SSEs. In any case, this logic places constraints on the frequency-magnitude distributions of slip events needed to balance the slip budget, as detailed below.

Fault slip can be inferred from surface deformation, measured using geodetic or remote sensing techniques, according to the theory of a dislocation embedded in an elastic medium (e.g., Cohen 1999, Okada 1985) as is customary for earthquakes. In the conceptual model of Figure 1, two key assumptions are made. The first is that, in the interseismic period separating episodic interplate slip events, the subduction interface is fully locked from the surface to a depth $z_d$. The second assumption is that the long-term slip rate on the interface is equal to the horizontal convergence rate across the plate boundary, with neither plate experiencing anelastic internal deformation. Although evidence for anelastic deformation of the overriding plate is commonly observed, it generally accounts for only a small fraction (a few percent) of the convergence rate (e.g., Briggs et al. 2008, Matsu’ura et al. 2014, Melnick et al. 2006, Saillard et al. 2009). Note that this second assumption can be strictly fulfilled only if the fault is planar. It is, however, a reasonable approximation to first order (Vergne et al. 2001).

The predicted surface displacements depend on the location of the downdip end of the locked fault zone and on the slip rate vector. The blue line in Figure 1b represents the theoretical surface...
Figure 1

Simple two-dimensional model of the seismic cycle on a subduction zone. The model assumes that, in the long run, plate convergence is entirely absorbed by slip along the subduction interface (the so-called megathrust). (a) Interseismic coupling. (b) Displacements at the surface. The blue line represents the theoretical surface displacement rate relative to the stable overriding plate. The dark yellow line represents the long-term surface displacement rate. The difference between the two curves, depicted by the red line, represents the contribution of transient slip events. (c) Displacements at depth, indicated by red arrows (coseismic) and blue arrows (cumulated over the interseismic period). The shallower portion of the megathrust, highlighted in red, is assumed to slip only during transient slip events (interplate earthquakes, afterslip, or slow slip events) and is fully locked during the interseismic period between these transients ($\chi_i = 1$ in panel a). The deeper part of the interface (beyond depth $z_d$), highlighted in blue, creeps at the plate convergence rate, $V$ ($\chi_i = 0$ in panel a).

displacement rate relative to the overriding plate. The dark yellow line represents the long-term surface displacement rate. The difference between the two curves, indicated by the red line, represents the long-term averaged contribution of transient slip events (the surface displacement resulting from all transient slip events summed over a long period and divided by the period duration $T$).
The partitioning of slip over the long-term average can be quantified using the following quantities:
- seismic coupling, $\chi_s$, the ratio of cumulative seismic slip to long-term slip;
- aseismic coupling, $\chi_a = \chi_{as} + \chi_{SSE}$, the ratio of cumulative aseismic transients (afterslip and SSEs) to long-term slip; and
- interseismic coupling, $\chi_i$, the ratio of the deficit of slip in the interseismic period to long-term slip, assumed to be stationary throughout the interseismic period.

All three terms range between 0 and 1 if we assume that slip occurs only in the direction of long-term motion on the fault. This formulation could be altered to allow for transient episodes of locking by adding one negative term representing transient backslip. The term coupling, in the tectonic geodesy context of this review, should not be confused with the coupling terminology commonly used to refer to stresses transmitted across a fault. Locally high coupling on a fault does not mean locally high stresses in the context of this article.

Closure of the slip budget (the condition that seismic slip and aseismic slip, whether due to SSEs or afterslip, sum to match the long-term slip at any point on the fault) gives

$$\chi_i = \chi_s + \chi_{as} + \chi_{SSE}. \quad (1)$$

The interseismic coupling, $\chi_i$, can be determined from modeling geodetic measurements of interseismic strain (e.g., Wang et al. 1995), provided the period does not encompass aseismic transients. The seismic coupling, $\chi_s$, can be estimated on the basis of the distribution of slip during identified interplate earthquakes (characterized by a thrust focal mechanism with one auxiliary plane parallel to the subduction interface) (e.g., Bird & Kagan 2004, Frohlich & Wetzel 2007, Tichelaar & Ruff 1993). This quantity is generally poorly constrained as the larger earthquakes, which contribute the most to fault slip, have return periods often much longer than the temporal coverage of earthquake catalogs (they might be missing, or their frequency may not be well constrained from these catalogs).

In the simple example of Figure 1, $\chi_i = 1$ over the locked portion of the subduction interface and $\chi_i = 0$ elsewhere. The rationale for this binary choice is that, if earthquakes are frictional slip instabilities, a given point on a fault should either stick or creep aseismically in the interseismic period. In fact, the creeping zone could slip at a rate either smaller or larger than the long-term slip rate due to stress redistribution during the seismic cycle. As stress builds up in the locked area during the interseismic period, the creeping zone at the edge is unloaded, resulting in a reduced slip rate there (Bürgmann et al. 2005, Hetland & Simons 2010). This effect is ignored in this simple reference model.

### 2.2. Relating Interseismic and Seismic Coupling

Earthquakes generally occur in a limited depth range, typically at depth less than 20 km within continents and between 10 and 50 km for subduction megathrusts (e.g., Ruff & Tichelaar 1996, Sibson 1982). A simple view would be that this seismogenic zone is the portion of the plate interface that is locked during the interseismic period (the white area, where $\chi_i = 1$, in Figure 2a). The deficit of seismic moment accumulating in the interseismic period due to locking of the plate interface is

$$\mathcal{M}_0 = \int_{\text{megathrust}} G V \chi_i \, dr,$$

where $G$ is the shear modulus (assuming for simplicity a homogeneous elastic halfspace) and $V$ is the long-term slip rate. In the simple binary case of Figure 2, this equation yields $\mathcal{M}_0 = GVLW$, where $W$ is the downdip width and $L$ is the along-strike length of the locked fault zone.
Figure 2

(a) A simple idealized view of the relationship between interseismic slip, $\chi$, and transient slip events on a megathrust. The interface is assumed to be fully locked ($\chi = 1$; white area) at depths between the updip limit, $z_u$, and the downdip limit, $z_d$, of the seismogenic zone. The upper (at depth between the trench depth, $z_t$, and $z_u$) and lower (at depth below $z_d$) portions are assumed to be creeping at the long-term slip rate $V (\chi = 0$; grey areas). Slip due to transient slip events within the locked portion (red patches) is assumed to sum to uniform slip on the long-term average. (b) Interplate earthquakes are assumed to obey the Gutenberg–Richter law (red lines). The ordinate shows, on a logarithmic scale, the number of events with moment magnitude ($M_w$) equal to or larger than a given value reported along the abscissa. The frequency of the maximum-magnitude earthquake (blue line) is determined by the slip budget closure condition (Equation 1), the $b$ value of the Gutenberg–Richter distribution (the slope of the red lines) and the ratio, $\alpha$, between aseismic transient slip and seismic slip (Equation 3).

Let $\alpha$ be the fraction of transient slip that is seismic [$\alpha = \chi_s / (\chi_s + \chi_{as} + \chi_{SSE})$], and let us now assume that the frequency-magnitude distribution of interplate earthquakes follows, on the long-term average, the Gutenberg–Richter law with exponent $b$. With these assumptions, it is possible to relate the rate of moment deficit accumulation to the return period of any earthquake with moment $M$ (Molnar 1979):

$$T(M) = \frac{1}{1 - 2b/3} \frac{M_{max}}{\alpha M_0} \left( \frac{M}{M_{max}} \right)^{2b/3},$$

where $M_{max}$ is the moment of the maximum-magnitude earthquake. This equation is valid only if $b < 3/2$; otherwise, the integral of cumulated moment over all moments is divergent (the integral does not diverge if the minimum moment cut-off is assumed non-null). The return period of the maximum-magnitude earthquake is then

$$T(M_{max}) = \frac{1}{(1 - 2b/3)\alpha} \frac{M_{max}}{M_0}. \quad (4)$$

The frequency $N_{max}$ of the earthquake with the maximum moment $M_{max}$ can then be expressed as a function of its moment magnitude [defined as $M_w = \frac{2}{3} \log(M_0) - 6$, where the moment is expressed in Newton meters and the logarithmic function is decimal],

$$\log N_{max} = -\frac{3}{2} M_w - 9 + \log M_0 + \log \left[ \alpha \left(1 - \frac{2b}{3} \right) \right]. \quad (5)$$

If interseismic strain is assumed to be released only by the return of a characteristic earthquake of moment $M_{char}$, the average return period is

$$T(M_{char}) = \frac{M_{char}}{M_0}. \quad (6)$$
corresponding to \( b = 0 \) and \( \alpha = 1 \) in Equation 4. The prefactor in Equation 4 is always larger than 1 [because \((1 - 2b/3)\alpha < 1\)], so the characteristic earthquake hypothesis tends to underestimate the return period of the assumed maximum-magnitude earthquake.

This simple model illustrates how the measurement of interseismic strain from geodesy can be used to place constraints on seismic hazard. We will see below, through the review of case examples and numerical simulations of fault dynamics, that some of the assumptions of the model do not seem to hold.

3. FAULT SLIP BUDGET: A REVIEW OF CASE EXAMPLES

3.1. The Himalayan Megathrust

The Himalayan range is a crustal-scale wedge formed of a series of imbricated thrust sheets that root into a single basal thrust fault, the Main Himalayan Thrust (MHT), which surfaces at the toe of the wedge (e.g., Hauck et al. 1998). The shortening rate of \( 21.0 \pm 1.5 \text{ mm/yr} \) inferred across the MHT from the deformation of Holocene river terraces (Lavé & Avouac 2000) is approximately equal to the geodetically inferred shortening rates of \( 20.5 \pm 1 \text{ mm/yr} \) in western Nepal and \( 17.8 \pm 0.5 \text{ mm/yr} \) in eastern Nepal (Figure 3) (Ader et al. 2012), suggesting that shortening across the Nepal Himalaya is nearly entirely the result of thrusting along the MHT. Inversion of geodetic data collected across the Himalaya over the past 20 yr (e.g., Ader et al. 2012, Bilham et al. 1997) indicates that the MHT is fully locked over a width of \( \sim 120 \text{ km} \) updip of the front of the High Himalaya, and north of that it is creeping (Figure 3). The transition to a fully creeping zone occurs over less than 30 km (Ader et al. 2012). Thus, the MHT seems to conform to first order to the simple model of the seismic cycle on a megathrust presented in the previous section.

Seismicity is localized along the downdip end of the locked fault zone, where the stress buildup rate is highest (Figure 3a) (e.g., Bollinger et al. 2004, Cattin & Avouac 2000). The locked fault zone has otherwise remained mostly silent beneath Nepal. Slip there must occur mostly during the largest earthquakes, such as the 1934 \( M_w 8.0 \) Bihar earthquake, as a result of coseismic slip due to the mainshock or of aftershocks and afterslip.

The relationship between the return period and the magnitude of the largest earthquake needed to close the slip budget can be estimated from Equation 4. If we assume that slip (\( \alpha = 1 \)) within the locked fault zone and a maximum possible magnitude of 8, the 1,000-km-long stretch of the Himalayan arc covered by the coupling model of Figure 3a would need to produce a \( M_w 8 \) earthquake approximately every 30 yr. In contrast, the \( M_w \geq 8 \) events (1950, 1934, 1505, and 1255) suggest a much longer return period. If we assume a maximum magnitude of 8.5, similar to the magnitude of the 1950 Assam earthquake, the return period of such an event would need to be approximately 300 yr, but the 1950 event is the only documented occurrence during the past 500 yr. In fact, the known seismicity over the past 500 yr accounts for less than 20\% of the slip deficit due to interseismic locking of the MHT (Ader et al. 2012). Extrapolating the frequency-magnitude distribution of instrumental seismicity to larger magnitudes based on the Gutenberg–Richter law also falls well short of accounting for the interseismic slip deficit (Figure 3b). The long-term seismicity rate would need to be approximately 10 times larger than the interseismic rate for the largest-magnitude earthquake to be \( M_w 8.5 \). That would mean, for example, that 9 out of 10 \( M_w 7 \) earthquakes along the Himalayan arc would be aftershocks of larger events. This discussion shows that, in this case, it is challenging to reconcile the geodetic data with the known instrumental and historical seismicity rate using the rationale proposed in the previous section.

We made the assumptions that (a) interseismic coupling measured over the past 20 yr can be extrapolated to longer timescales, (b) aseismic slip within the locked fault zone does not contribute
(a) Relocated seismicity in the Nepal Himalaya and pattern of interseismic coupling, $\chi_i$, on the Main Himalayan Thrust over the period from 1995 to 2010. The inset illustrates the location of the study area. The yellow ellipse shows the approximate location of the 1934 $M_w$ 8.0 Bihar earthquake. (b) Gutenberg–Richter plot of the seismicity in Nepal, using the National Seismic Centre (NSC) local catalog (1995–2001), the GCMT catalog (1976–2010), and the historical catalog of Ambroseys & Douglas (2004) over the preceding 500 yr for $M_w > 8$ earthquakes and the preceding 200 yr for $M_w > 7.5$ earthquakes. The red lines are the distribution that the seismicity would follow if 100% of the moment deficit was released seismically ($\alpha = 1$) following a Gutenberg–Richter distribution with $b = 1$, up to maximum magnitudes of 8, 9, and 10. The blue line shows the return period of the maximum-magnitude earthquake needed to close the slip budget for $\alpha = 1$. (c) Comparison between interseismic coupling and temperature along a section across the Himalaya at the longitude of Kathmandu. Figure modified with permission from Ader et al. (2012).

Any of these assumptions might be invalid. For example, the locked fault zone might produce spontaneous SSEs or slip as a result of afterslip following large earthquakes, implying $\alpha < 1$. In order to close the slip budget for a maximum magnitude of 8.5, $\alpha$ would need to be on the order of 0.1, an improbably low value. There are only a few examples of documented afterslip following...
large continental thrust earthquakes (Hsu et al. 2002, 2009a; Huang et al. 2014; Jouanne et al. 2011). In these examples, afterslip occurred downdip of the seismic ruptures, presumably within the transition zone between the fully locked and fully creeping zone rather than within the locked fault zone. Additionally, no significant aseismic transient has been detected in the Nepal Himalaya over the 20 yr covered by geodetic data. It is therefore improbable that aseismic slip contributes enough slip to help close the slip budget. The largest possible earthquake in the Nepal Himalaya must exceed $M_w 8.5$ and/or be more frequent than the historical data suggest ($T < 300$ yr). In any case, it is probably incorrect to infer the magnitude and frequency of the largest event based on the frequency-magnitude distribution of the smaller earthquakes recorded in the interseismic period.

Interseismic coupling on the MHT is observed to vary essentially in the downdip direction: The fault is essentially fully locked south of the front of the Higher Himalaya and fully creeping north of it. The transition from unstable sliding to stable sliding (Figure 3c) occurs where the temperature is estimated to increase from approximately 300°C to 500°C (Ader et al. 2012, Avouac 2003), consistent with the common view that temperature is the dominant controlling factor (e.g., Ader et al. 2012, Avouac 2003, Blanpied et al. 1991, Sibson 1982). Other factors (pressure, metamorphic grade, and fluid contents) also vary primarily in the downdip direction and may also contribute to this transition.

3.2. The Longitudinal Valley Fault

The Longitudinal Valley Fault (LVF) in Taiwan (Figure 4) marks the suture zone along which the Luzon volcanic arc collided with the continental margin of southern China. This fault accommodates about half of the 93 mm/yr convergence between the Philippine Sea and Eurasian plates and dips steeply eastward by approximately 60°. The LVF produced two $M_w 7.3$ earthquakes and one $M_w 7.1$ earthquake in 1951 (Shyu et al. 2007) and, more recently, the $M_w 6.8$ Chengkung earthquake in 2003 (Hsu et al. 2009b, Mozziconacci et al. 2009). Surface displacements over the past 15 yr are well documented thanks to the Taiwan Earthquake Center’s network of continuous global positioning system (GPS) stations (http://gps.earth.sinica.edu.tw/), campaign-mode GPS measurements (Yu & Kuo 2001), leveling data (Chen et al. 2012, Ching et al. 2011), and interferometric synthetic aperture radar (InSAR) measurements (e.g., Champenois et al. 2012). These data reveal that the fault is creeping at the surface all along its southern half but is locked along its northern half. The 2003 Chengkung earthquake ruptured an isolated locked patch, at a depth between 7 and 20 km and embedded in a creeping zone (Figure 5a) (Thomas et al. 2014a). The patch slipped by about 1 m during the earthquake and locked again immediately after it. The area surrounding the rupture area was creeping aseismically before the quake at about half the long-term slip rate (Figure 5b). The creep rate increased suddenly at the time of the earthquake and returned to the preseismic rate within a few years (Figure 5d). The overlap between distributions of seismic and aseismic slip is more significant than expected from the smearing effect of the limited resolution of the inversion (Thomas et al. 2014a).

Figure 6 shows the frequency-magnitude distribution of seismicity in the area of the Chengkung earthquake over 20 yr compared with the frequency-magnitude distribution of the largest earthquake needed to release interseismic strain, calculated using the interseismic coupling model. A closed slip budget is obtained for a maximum-magnitude earthquake with moment magnitude of $M_w 6.8$ and 34-yr return period, remarkably consistent with extrapolation of the instrumental seismicity rate based on the Gutenberg–Richter law. This consistency probably appears because the seismicity and geodetic data cover a time span comparable to the return period of the largest event and include such an event, unlike the case of the Himalayan arc.
Figure 4

(a) Plate tectonic setting of the Longitudinal Valley Fault (LVF) in Taiwan (inset shows location) and surface displacements measured between 2007 and 2010 (Champenois et al. 2012, Thomas et al. 2014a). The color coding shows the amplitude of the mean line of sight velocity measured from the persistent scatterer technique applied to ALOS PALSAR synthetic aperture radar images. The black star shows the epicenter of the 2003 $M_w$ 6.8 Chengkung earthquake. Black arrows show the ascending track direction and the line of sight, which has an incidence of about 35° relative to vertical. The box denotes the location illustrated in Figure 5. (b) Simplified geological map of eastern Taiwan. Formations east of the LVF include Mio-Pliocene volcanics (Tuluanshan), Plio-Pleistocene forearc and intra-arc sediments (Takangkou), and a Pliocene mélange (Lichi). The formation west of the LVF consists of quaternary deposits. Figure modified with permission from Thomas et al. (2014a).
Another major difference between the Himalayan and LVF examples is that on the LVF, aseismic slip accounts for approximately 85% of the long-term slip rate within the 0–26-km seismogenic depth range (Thomas et al. 2014a); in contrast, aseismic creep within the seismogenic zone is likely negligible in the Himalayan case. This is because, whereas the Himalayan example shows
Figure 6
Seismic and aseismic slip budget closure test on the southern portion of the Longitudinal Valley Fault (in the area of the 2003 $M_w$ 6.8 Chengkung earthquake) based on the frequency-magnitude distribution of seismicity from 1991 to 2010. The seismicity follows the Gutenberg–Richter law (red line), yielding a $b$ value of 0.95 for events with magnitudes above the magnitude of completion ($M_c = 2$). The blue line shows the return period of the maximum-magnitude earthquake when its magnitude is varied so that the interseismic deficit slip is balanced by the seismic slip resulting from all the events defining the Gutenberg–Richter distribution (Equation 4). The star at the intersection of the blue and red lines indicates the maximum magnitude ($M_w$ 6.8) and predicted return period (34 yr) of a maximum-magnitude earthquake determined from the slip budget closure.

only downdip variation of interseismic coupling, the LVF example shows strong along-strike variations that cannot be the result of a temperature control of the mode of slip.

The correlation between surface creep and the Lichi mélange (Figure 6) suggests that this lithology favors aseismic creep (Thomas et al. 2014b). This formation, characterized by a strongly foliated clay matrix with up to kilometer-sized exotic blocks, has been interpreted as either an olistostrome or a collisional mélange (e.g., Huang et al. 2008, Page & Suppe 1981), but in either case...
case it is similar to an un lithified subduction mélangé. Thus, the oceanic origin of the LVF may explain the prevalence of aseismic creep on the LVF.

### 3.3. The Sumatra Megathrust

The pattern of interseismic coupling and the history of large earthquakes on the subduction zone offshore Sumatra are now relatively well documented (e.g., Briggs et al. 2006; Chlieh et al. 2008; Konca et al. 2008a; Meltzner et al. 2010, 2012; Natawidjaja et al. 2000; Philibosian et al. 2014; Prawirodirdjo et al. 2010; Sieh et al. 2008). Figure 7 shows the pattern of interseismic coupling derived from the joint inversion of geodetic strain measured from GPS and coastal uplift measured from coral growth from 1950 to 2000 (Chlieh et al. 2008). The model assumes no coupling at the trench, where the resolution of the inversion is poor. A key and robust feature of the model is the presence of large along-strike variations in coupling. Aseismic creep, integrated over all parts of the megathrust within the 0–40-km seismogenic depth range, accounts for 14% to 43% of the long-term slip rate. This corresponds to a mean interseismic coupling of 0.57 to 0.86 at depths less than 40 km. The rupture areas of large interplate earthquakes ($M_w$ 8.4 and 7.9 in 2007, $M_w$ 8.6 in 2005, $M_w$ $\geq$ 8.7 in 1797, and $M_w$ $\geq$ 8.8 in 1833) correlate relatively well with regions of high interseismic coupling (Figure 7). Seismic slip related to large earthquakes has accounted for only 21% to 41% of megathrust slip at $<40$ km depth over the past 300 yr. These events have also triggered aseismic slip and aftershocks that need to be accounted for in the slip budget.

Figure 8 shows the spatial distribution and temporal evolution of a fterslip over the course of a year following the 2005 $M_w$ 8.6 and 2007 $M_w$ 8.4 earthquakes (Hsu et al. 2006, Konca et al. 2008b). In both cases, a fterslip was aseismic, as it released a cumulative moment equal to 20–30% of the coseismic moment, whereas aftershocks released 10 times less cumulative moment (<3% of the coseismic moment). Afterslip occurred mostly updip and downdip of the rupture areas, in the zone where interseismic coupling is tapering off, suggesting that creep there is enhanced by the stress increase when the locked areas rupture. Notably, the area north of the 2007 $M_w$ 8.4 rupture area did not produce much a fterslip (Figure 8e) until an earthquake swarm began with a $M_w$ 5.4 event on February 19, 2008 (Figure 9). Inversion of the GPS data (corrected for detectable coseismic displacements) indicates that this area unlocked and produced a fterslip with a cumulative moment 3 times as large as that released by the seismic swarm (equivalent to a $M_w$ 7.0 event) (Figure 9). It seems that the swarm activity controlled the timing of aseismic slip, contrary to the cases in which aftershocks seem to be driven by a fterslip (e.g., Hsu et al. 2006, Perfettini & Avouac 2004a, Savage et al. 2007, Schaff et al. 1998, Yu et al. 2013). Regions of swarm activity identified elsewhere on the Sumatra megathrust are also observed to coincide with creeping areas (Holtkamp & Brudzinski 2014). The Mentawai region additionally produced a large tsunamigenic $M_w$ 7.8 earthquake in...
Interseismic coupling, $\chi_0$

0 0.2 0.4 0.6 0.8 1.0

EQ

Andaman Islands

INDIAN PLATE

SUNDA PLATE

AUSTRALIAN PLATE

Wharton Ridge

Investigator Fracture Zone

Mw 8.6 2004

Mw ~8.7 1797

Mw ~8.8 1833

Mw 7.9 2007

Mw 8.4 2007

Padang

Bengkulu

V = 5.7 cm/yr

Latitude

Longitude

Annu. Rev. Earth Planet. Sci. 2015.43:239-271. Downloaded from www.annualreviews.org on 01/22/18. For personal use only.
2010 (Figure 8) (Hill et al. 2012, Lay et al. 2011b). A similar tsunamigenic $M_w$ 7.6 earthquake occurred north of Padang in 1907 and is estimated to have ruptured the shallow portion of the megathrust updip of the 2005 $M_w$ 8.6 earthquake (Kanamori et al. 2010).

The Sumatra example shows that a locked patch can rupture in a variety of ways. The patch that ruptured in 2005 had ruptured in a similar event in 1861, and the time span between the two events is approximately consistent with the time needed to build up the elastic strain released in 2005 (Chlieh et al. 2008). The patch beneath the Mentawai Islands, offshore Padang (Figure 7), shows a more diverse mode of rupture. It produced $M_w > 8.7$ earthquakes in 1797 and 1833, with apparently overlapping rupture areas (Figure 7), and the southern portion ruptured again in
2007 but during a sequence of smaller events ($M_w \leq 8.4$). The 2007 $M_w 8.4$ and 7.9 earthquakes ruptured approximately the same total area as in 1833 but did not produce much slip where the 1797 and 1833 ruptures overlap, possibly because of the stress drop there associated with these previous ruptures. Although large earthquakes have not been observed to propagate through zones of aseismic creep in Sumatra, these zones have produced aftershocks, earthquake swarms, and tsunamigenic earthquakes.

It is possible to test crudely whether seismic slip and aseismic slip on the Sumatra megathrust add up to the long-term slip imposed by plate convergence according to the principle outlined in the previous section. We assume that of the $4 \times 10^{20}$ N m yr of moment deficit accumulating in the interseismic period, 75% is released by seismic slip and 25% by aseismic transients. Using Equation 4, and given the $b$ value of the frequency-magnitude distribution of seismicity, we can estimate the frequency of the largest earthquake needed to close the slip budget (blue line in Figure 10). A $M_w 8.5$ event would need to occur every 56 yr on average (Natawidjaja et al. 2000, Newcomb & McCann 1987, Philibosian et al. 2014). The return period of a $M_w 9.0$ earthquake would need to be about 316 yr. Only the 1797 event is estimated to have possibly reached such a magnitude (Philibosian et al. 2014), but four events over the past 300 yr probably exceeded $M_w 8.5$ (Natawidjaja et al. 2000, Newcomb & McCann 1987, Philibosian et al. 2014). Extrapolation of the frequency-magnitude distribution of seismicity recorded before 2005 would imply a $M_w > 9$ earthquake with a $> 1,000$-yr return period, but this extrapolation is not valid, as the aftershocks of
Moment magnitude, $M_w$

Number of earthquakes per year with moment magnitude $> M_w$

Return period

Before the 2005 $M_w 8.6$ earthquake

Whole period

Figure 10

Seismic and aseismic slip budget closure test on the Sumatra megathrust based on the frequency-magnitude distribution of seismicity from 1976 to 2014 (GCMT catalog). The seismicity before the 2005 $M_w 8.6$ earthquake follows the Gutenberg–Richter law (red triangles), yielding a $b$ value of 0.94 for events with magnitudes above the magnitude of completion ($M_w \geq 5$). The seismicity over the whole period follows the Gutenberg–Richter law (orange circles), yielding a $b$ value of 0.78. The blue line shows the return period of the maximum-magnitude earthquake when its magnitude is varied so that the interseismic deficit slip is balanced by the seismic slip resulting from all the events defining the Gutenberg–Richter distribution (Equation 4), assuming a $b$ value of 0.9 and seismic slip to account for 75% of interseismic deficit of slip ($\alpha = 0.75$). The star indicates the rate of $M_w > 8.5$ earthquakes since 1700 based on historical and paleogeodetic studies (Natawidjaja et al. 2000, Newcomb & McCann 1987, Philibosian et al. 2014).

Thus, historical seismicity seems to be representative of the long-term seismicity required to close the budget of seismic and aseismic slip in Sumatra. Note that this test, at the scale of the whole study area (Figure 7), does not mean that the slip budget closes locally. Despite the approximate balance of strain buildup and release over the past 300 yr, it is clear that large elastic strain must have built up offshore Padang since it last ruptured in 1797. Although the slip-budget analysis could be used to evaluate the probability of such a rupture in the future, it would be better to rely on a physical model that takes into account the spatiotemporal evolution of strain buildup and release.

3.4. Seismic and Aseismic Slip on Other Subduction Megathrusts

It is useful to compare the results on the Sumatra megathrust with results obtained from the study of other subduction megathrusts. First, spatially heterogeneous interseismic coupling seems to be
the rule rather than the exception (e.g., Chlieh et al. 2011; Cross & Freymueller 2007; Fournier & Freymueller 2007; Hashimoto et al. 2009; Loveless & Meade 2010, 2011; Manaker et al. 2008; M´etois et al. 2012, 2014; Moreno et al. 2011; Radiguet et al. 2012; Wallace et al. 2004, 2009). Second, seismic ruptures tend to be confined within locked patches, as observed on the Sumatra megathrust and the LVF. A particularly convincing example is the 2012 $M_w$ 7.6 earthquake in Costa Rica, which ruptured a well-documented locked patch beneath the Nicoya Peninsula (Feng et al. 2012, Protti et al. 2014, Yue et al. 2013). A corollary is that large megathrust ruptures tend to arrest systematically when they enter areas of interseismic aseismic creep.

Most studies of postseismic deformation following large megathrust ruptures show that near-field deformation is due to aseismic afterslip (e.g., Baba et al. 2006; Chlieh et al. 2004, 2007; Hutton et al. 2001; Lin et al. 2013; Melbourne et al. 2002; Miyazaki et al. 2004; Ozawa et al. 2012; Perfettini et al. 2010). Here I do not discuss the effect of viscoelastic relaxation. Readers interested in this topic are referred to Wang et al. (2012a); the effect should in principle induce long-term variations of the fault loading rate, which I ignore here for simplicity. For $M_w > 7.5$ events, the moment released by afterslip is generally 10–40% of the coseismic moment, possibly with a trend toward a larger proportion of aseismic creep for larger earthquakes (Lin et al. 2013). Afterslip is most often observed downdip of the rupture area of those earthquakes, near the downdip edge of the zone of strong interseismic locking. Afterslip at shallower depth, where the sensitivity and resolution of geodetic inversions are poorer, is less common but has been observed on occasion (and not only updip of the ruptured patch, as observed following the 2005 and 2007 Sumatra earthquakes). For example, afterslip following the 2007 $M_w$ 8.0 Pisco earthquake offshore southern Peru occurred within different patches complementing the area of coseismic slip (Perfettini et al. 2010). Afterslip following the Maule earthquake occurred on patches that overlap with the seismic rupture area and with the interseismically locked zone (Bedford et al. 2013).

There are numerous examples showing that, like in Sumatra, a single segment of a megathrust sometimes ruptures in $M_w > 8.5$ earthquakes and other times ruptures in sequences of smaller events (e.g., Kanamori 1986, Kanamori & McNally 1982, Lay et al. 1989). Examples of repeating similar megathrust earthquakes are rare; it is more common to see complementarity (e.g., Park & Mori 2007) or diversity when successive ruptures are compared (e.g., Thatcher 1990). The Tohoku-oki earthquake of 2011 is interesting in that regard. It initiated in an area of high interseismic coupling and ruptured areas that had ruptured in separate $M_w$ 7.5 earthquakes in the past (Hashimoto et al. 2009, Loveless & Meade 2011). In contrast to the megathrust earthquakes discussed above, this event propagated all the way to the trench (Fujiwara et al. 2011, Lay et al. 2011a). Given the lack of near-trench resolution of geodetic inversions, this observation led to a discussion of whether the shallow portion of the megathrust is locked or creeps during the interseismic period. Although the matter remains unresolved, there are hints that this area can both creep in the interseismic and postseismic periods and slip during large earthquakes. If the megathrust were locked all the way to the trench, closure of the slip budget would require a return period for $M_w > 9.0$ events of only a few centuries (Avouac 2011, Perfettini & Avouac 2014), and the 2011 Tohoku-oki earthquake was the only such event since the 869 Jogan earthquake (Minoura et al. 2001, Sugawara et al. 2012). It is therefore more probable that the shallow portion of the megathrust there is mostly creeping in the interseismic period but can rupture occasionally, possibly as a result of strong dynamic weakening (Noda & Lapusta 2013).

Subduction zones and some continental faults can produce aseismic slip transients that are not necessarily triggered by nearby earthquakes (e.g., Douglas et al. 2005; Drage et al. 2004; Hirose et al. 2010; Lowry et al. 2001; Miller et al. 2002; Peng & Gomberg 2010; Radiguet et al. 2011, 2012; Schwartz & Pekosky 2007; Wallace & Beavan 2010; Yoshioka et al. 2004; Zigone et al. 2012). These aseismic transients tend to occur around the downdip end of the seismogenic zone.
and are associated with tremors, suggesting the involvement of fluids. This inference is supported by geophysical evidence for very low $V_p/V_S$, for example in the area of SSEs offshore Guerrero in Mexico (Song et al. 2009). Another example showing the close association of aseismic transients with earthquake swarms, and probably transient fluid flow, is the South America megathrust offshore southern Ecuador (Chlieh et al. 2014).

4. DISCUSSION OF OBSERVATIONAL CONSTRAINTS ON SEISMIC AND ASEISMIC SLIP

4.1. Stationarity of Interseismic Coupling and Loading Rates

In Equation 1, we assumed that interseismic coupling, $\chi_i$, is constant in time. This hypothesis can be tested only if continuous geodetic time series are available and might not hold when extrapolated over a longer time period. For example, continuous GPS measurements collected over >15 yr in the Nepal Himalaya show a temporally uniform pattern of interseismic strain. However, it might not be correct to assume that the pattern is stationary at the timescale of the centuries-long periods separating the largest earthquakes. There is geodetic and paleogeodetic evidence that the pattern of coupling on portions of the Sumatra megathrust changed through time (Philibosian et al. 2014, Prawirodirdjo et al. 2010). In particular, the southern segment, which is currently freely creeping, was apparently more locked before 2003. The change coincided with a $M_w$ 8.0 intraslab earthquake (Abercrombie et al. 2003). Similarly, interseismic strain around the rupture area of the $M_w$ 9.2 1954 Alaska earthquake shows quite significant temporal variations due to afterslip and transient aseismic events (i.e., SSEs) (Ohta et al. 2006). Locking of the Hikurangi megathrust also shows considerable variations due to large-scale SSEs around the downdip edge of the locked zone. In these examples, interseismic coupling can be distinguished from these slow events because the available geodetic time series cover a time period much longer than the duration of the aseismic transients. However, aseismic transients can occur over a wide range of timescales (from hours to several years) and can have very long return periods. This makes the determination of $\chi_i$ and $\chi_s$ from geodetic observations challenging.

I have also assumed that the loading rate, represented by the long-term slip rate $V$ in Equation 2, can be considered constant. This may not be true in general (Bennett et al. 2004, Friedrich et al. 2003, Peltzer et al. 2001). Viscoelastic relaxation or fault interactions could result in temporal variations at the earthquake-cycle timescale (Johnson & Segall 2004, Perfettini & Avouac 2004b).

4.2. The Influence of Lithology, Temperature, Roughness, and Fluids on Fault Creep

The Himalayan example conforms to the view that the ability of a fault to stick or creep in the interseismic period varies only along dip and that the transition to stable creep is probably thermally controlled and occurs at a depth and location consistent with the critical temperature of approximately 350°C to enable steady frictional slip for quartzofeldspathic rocks (e.g., Blanpied et al. 1991). In this case, the seismogenic zone seems to coincide approximately with the locked fault zone. If this rationale were to apply in general, the coupled seismic thickness (Bird & Kagan 2004), which is defined as

$$Z_s = \frac{\int_{\text{fault area}} \chi_s \, dx}{L},$$

(7)

where $L$ is the along-strike length of the fault, would in general be comparable to the width of the seismogenic zone and of the locked fault zone as well. Bird & Kagan (2004) carried out a global
analysis of $Z_s$ on major faults based on the summation of seismic moments and fault slip rates. They found that $Z_s$ is $\sim 8$ km for continental strike-slip faults and $\sim 18$ km for thrust faults (which converts to a vertical extent of 9 km for a standard dip angle of 30°). Given that the seismogenic zone generally extends to $\sim 15$ km depth within continents (Sibson 1985), seismic slip seems to account for only a fraction of the long-term slip within the seismogenic zone, suggesting that a significant portion of slip there is aseismic. A few examples of continental faults creeping at the surface have long been known, such as the creeping segment of the San Andreas Fault in California (de Michele et al. 2011, Titus et al. 2005, Tocher 1960), the Hayward Fault in California (Bonilla 1966, Cluff & Steinbrugge 1966), and the North Anatolian Fault in Turkey (Ambraseys 1970, Cakir et al. 2005, Kaneko et al. 2013). In addition to the LVF, reviewed above, examples of faults creeping at shallow depth have now accumulated thanks to GPS geodesy and InSAR (e.g., Donnellan et al. 2014; Doubre & Peltzer 2007; Duquesnoy et al. 1994; Hreinsdottir & Bennett 2009; Huang et al. 2009; Jolivet et al. 2012, 2013; Kaneko et al. 2013; Lohman & McGuire 2007; Sabadini et al. 2009; Shirzaei et al. 2013; Sylvestre et al. 1993; Szeliga et al. 2012; Wei et al. 2011).

Creep seems to be common on oceanic transform faults and normal faults at oceanic spreading centers: Bird & Kagan (2004) found $Z_s$ to be $<1$ km on these types of faults, and McGuire et al. (2005) found that 85% of slip on transform faults along the East Pacific rise was aseismic (McGuire et al. 2005). Bird & Kagan (2004) estimated seismic thickness on megathrusts to be approximately 18 km, well short of the $>40$-km vertical extent of the seismogenic zone and corresponding to a seismic coupling of approximately 0.45 when averaged over the 0–40-km depth range. This value is consistent with the interseismic coupling inferred for megathrusts (typically 0.5–0.8). One frequent source of uncertainty in the estimated locking distributions on oceanic megathrusts is the lack of geodetic constraints near the trench (Loveless & Meade 2011). Most studies assume zero interseismic coupling at the trench due to the observation that seismicity generally cuts off at $<10$ km depth and the expectation that clay-rich sediment dragged down the megathrust should favor aseismic creep (e.g., Byrne et al. 1988, Hyndman et al. 1997). The observation of aseismic creep near the trench following the 2005 and 2007 earthquakes offshore Sumatra, and possibly following the 2011 Tohoku-oki earthquake (Perfettini & Avouac 2014), supports this hypothesis. The sea-bottom geodetic data from Japan and Peru (Gagnon et al. 2005, Matsumoto et al. 2008), the only two places where such data are currently available, suggest weak near-trench coupling as well. However, the inference of seismic slip to the trench during the Tohoku-oki earthquake (Fujwara et al. 2011, Lay et al. 2011a) and aftershocks (Kiser & Ishii 2013) challenges the view that the mode of slip on a fault is a permanent feature.

Fault patches that creep interseismically should generally inhibit seismic rupture propagation. There is, however, evidence that slip on a fault can sometimes switch from seismic to aseismic, possibly back and forth. Microstructural evidence from exhumed subduction megathrusts suggests that the same shear zone can host both seismic and aseismic deformation (e.g., Rowe et al. 2012b). There are also geodetic and SAR observations suggesting this possibility: The observation of aseismic creep on some portions of the fault that ruptured during the 1999 Izmit earthquake (Cakir et al. 2003) and on a portion of the rupture area of the 1964 subduction earthquake in Alaska (Zweck et al. 2002) shows that seismic ruptures can occasionally propagate into areas that creep in the interseismic or postseismic period.

The observation that seismic coupling is small ($\sim 0.1$) on oceanic transform faults with slip rates of $>4$ cm/yr but large (0.9) on faults slipping more slowly (Bird & Kagan 2004) suggests that the slip mode also depends on the loading rate. This dependence could be due to material properties, or possibly to the effect of fluid pressure resulting from the coupling between fault-zone deformation and fluid circulation. A transition from stable sliding to stick-slip as the loading
rate is increased has been observed for some rock types in laboratory experiments (e.g., Niemeijer & Spiers 2006).

The presence of serpentinite has long been suspected to play a key role in explaining the low strength and creeping behavior of megathrusts, oceanic faults, and some continental faults (DeShon et al. 2006, Hilairet et al. 2007, Hyndman et al. 1997, Irwin & Barnes 1975, Moore et al. 1997, Reinen et al. 1991). Some studies have also pointed to the possible central role of other minerals, such as talc or saponite (Moore & Rymer 2007, 2012). However, a review of the literature suggests it is unlikely that the presence of a particular mineral is the critical factor (Graymer et al. 2005; Hadizadeh et al. 2012; Holdsworth et al. 2011; Moore & Rymer 2007, 2012; Thomas et al. 2014b). In fact, it might take only some proportion of clay to promote creep, as suggested by the example of the LVF (Thomas et al. 2014a,b). Clay might also explain low near-trench coupling on subduction zones (Byrne et al. 1988). The clay-rich and poorly lithified sediments being accreted at the southern tip of Taiwan also show widespread and rapid aseismic deformation (Fruneau et al. 2001, Huang et al. 2009). Creep is also probably favored by hydrothermal alteration (Donnellan et al. 2014, Doubre & Peltzer 2007, Duquesnoy et al. 1994). Most clays, independent of their particular composition, are indeed observed to promote steady creep in laboratory experiments (den Hartog et al. 2012, Ikari et al. 2009, Saffer et al. 2012, Saffer & Marone 2003), and stable creep prevails in laboratory experiments on natural or synthetic clay-rich gouges (e.g., Niemeijer et al. 2008, Niemeijer & Vissers 2014).

Fault creep could result from the intrinsic mechanical properties of the fault-zone rocks, as observed in the laboratory, but could also result from other factors correlated with lithology, in particular fluid content. Fluids might promote pressure solution creep and hence aseismic creep, depending on their availability and chemistry. Microstructural studies of fault-zone rock samples deformed in laboratory experiments suggest that aseismic creep seems to be associated with pressure solution creep (Gratier et al. 2011, Niemeijer et al. 2008, Rowe et al. 2012b). Pressure solution creep can be strongly enhanced in the presence of CO$_2$ (Le Guen et al. 2007, Liteanu et al. 2012) and can therefore be a major creep mechanism in systems flushed with CO$_2$ of metamorphic or volcanic origin. Pressure solution is enhanced when soluble minerals (e.g., quartz and feldspar) come into contact with insoluble minerals (e.g., phyllosilicates) (Gratier et al. 2013), so tectonic mixing within a fault gouge or a subduction mélangé might promote aseismic creep.

Fault creep could also be mechanically promoted by elevated pore pressure. The reduction of the effective normal stress can indeed bring a potentially seismic rate-weakening fault into the conditionally stable domain favoring aseismic creep (e.g., Scholz 1998). This possible role of fluids in promoting aseismic creep is supported by the observation of aseismic slip–triggered fluid injection experiments in drill holes (Bourouis & Bernard 2007, Calo et al. 2011, Cornet et al. 1997, Scotti & Cornet 1994). Evidence for elevated pore-fluid pressure along aseismic portions of subduction zones based on geophysical observations (low $V_p/V_S$) (Moreno et al. 2014) or wedge dynamics (e.g., Cubas et al. 2013) provides another indication that fluids promote aseismic creep.

This notion could apply to the Himalaya, where the transition from the locked fault zone to the creeping zone correlates with a zone of high conductivity suggesting a well-connected fluid phase (Lemonnier et al. 1999). These fluids, which are likely sourced from metamorphic dehydration of the underthrusting Indian plate, could promote creep by reducing effective normal stress or enhancing aseismic deformation mechanisms such as pressure solution creep. Metamorphic fluids might play a similar role in oceanic subduction zones (Saffer & Tobin 2011), especially at the downdip end of the seismogenic zone, where numerous observations, such as tremors (e.g., Obara et al. 2004, Shelly et al. 2007) and tide-correlated slip events (Hawthorne & Rubin 2010), suggest the presence of fluids with high pore pressure.
Fault roughness may also influence fault slip mode: Wang & Bilek (2014) observed that a rough seafloor seems to promote aseismic creep on a subduction megathrust, possibly due to stress heterogeneities resulting from the geometric irregularities of the plate interface. The zone of low coupling around the equator on the Sumatra megathrust (Figure 7) correlates with the subduction of the Investigator Fracture Zone, and there are examples of similar correlations elsewhere [e.g., the Nazca Ridge in southern Peru (Perfettini et al. 2010)]. However, the zone of low coupling south of Bengkulu in Sumatra does not correlate with a particularly rough seafloor, and the Wharton Ridge is subducted beneath a strongly locked patch. Similarly, many studies have argued for a correlation or anticorrelation between the locations of large subduction earthquakes and subducting seamounts (e.g., Mochizuki et al. 2008, Watts et al. 2010, Yang et al. 2012). The correlation between fault roughness and interseismic coupling could also result from mineralogical effects: For example, hydrothermal alteration favored by faulting could cause serpentinization of the seafloor around the Investigator Fracture Zone, which could directly lubricate the megathrust or indirectly lubricate it through fluid release during subduction.

4.3. Persistent Locked Patches but Noncharacteristic Seismic Ruptures

The observations reviewed above suggest that patterns of interseismic coupling on faults probably persist throughout the interseismic period. Measurement of interseismic strain, and the associated pattern of coupling, can therefore in principle reveal the patches where stress is building up to be released in future seismic ruptures. This information might be used to place constraints on the seismic potential of a given fault. To do so, some assumptions are needed, as there is no unique solution. The repetition of a characteristic earthquake is not observed, as shown in the examples reviewed here and elsewhere (e.g., Schwartz 1999). Close inspections of examples of repeating events at the same location, such as the repeating \( M_w \) 4.9 earthquakes on the Japan megathrust offshore Kamaishi (e.g., Uchida et al. 2012) and the repeating \( M_w \) 6 earthquakes on the Parkfield segment of the San Andreas Fault (e.g., Bakun et al. 2005, Segall 1993), show that the largest events generally differ in their details. Successive ruptures might in fact complement each other (e.g., Custódio & Archuleta 2007, Park & Mori 2007) so that, in the long run, the fault slips everywhere at the same slip rate, as pictured in Figure 2. However, some characteristics, probably imparted by interseismic coupling patterns, can lead to systematics involving the repetition of quite similar larger earthquakes that could be considered to represent characteristic events. The smaller events probably create a different stress distribution from one cycle to another, so that successive larger events, however similar, always differ in their details. Assuming that the earthquakes follow the Gutenberg–Richter distribution provides a way to constrain the characteristics of the larger earthquakes. As we have seen, this approach is difficult to apply when the seismicity data cover only a fraction of the interseismic period, as in the Himalayan example. For these reasons, a more physical approach based on dynamic modeling of the seismic cycle would be appropriate. The promises and pitfalls of such an approach are discussed in the next section.

4.4. Constraints on Fault Rheology from Time Evolution of Afterslip

The time evolution of afterslip can be used to constrain fault rheology (e.g., Marone et al. 1991, Montési 2004). A now well-established feature is that afterslip, \( S \), increases approximately as a logarithmic function of time, \( t \) (e.g., Hearn et al. 2002; Hsu et al. 2002, 2006; Hutton et al. 2001; Marone et al. 1991; Melbourne et al. 2002; Perfettini et al. 2005; Savage et al. 2005):

\[
S = S_0 \log \left( \frac{t + t_c}{t_r} \right),
\]
where \( t_r \) is the characteristic time of relaxation. The critical time, \( t_c \), introduced in this equation to avoid the singularity at \( t = 0 \), is generally not resolvable from observation.

This time evolution is similar to the time evolution of transient creep in rocks and metals, confirming Benioff’s (1951) intuition that postseismic relaxation is governed by a creep recovery process (Savage et al. 2005). It is doubtful, however, that the mechanism governing afterslip in nature is the same as the one governing transient creep in rock samples in the laboratory. A variety of mechanisms could lead to a similar phenomenology in nature due to the scale difference and the possible role of fluids. For example, Gratier et al. (2014) observed that pressure solution creep, a mechanism that in principle obeys a Newtonian rheology in steady state, can lead to very nonlinear behavior resembling transient creep and postseismic deformation in response to a sudden fracturing event (because the fractures temporarily facilitate mass transfer and then heal with time). In any case, the \(~1/t\) Omori law of aftershock decay (e.g., Kisslinger 1996) may well be driven by afterslip (Hsu et al. 2006, Kato 2007, Perfettini & Avouac 2004a, Perfettini et al. 2010, Yu et al. 2013).

If we assume that the stress evolution during afterslip is dominated by elastic unloading and that the geometry of the slipping zone is not changing with time, fault friction is a linear function of slip,

\[
\mu = \mu_0 - kS,
\]

with \( k \) being the stiffness associated with the sliding area embedded in the elastic medium. If we now assume that friction depends only on the sliding rate, neglect the critical time introduced in Equation 8, derive Equation 8 with time to express time as a function of slip rate, and substitute into Equation 9, we get

\[
\mu = \mu_0 + kS_0 \log \frac{S}{S_0},
\]

where \( \dot{S} \) is the sliding velocity and \( S_0 = S_0/t_r \) is a reference slip rate. The logarithmic dependency of friction on the sliding rate can thus be derived from the observation that afterslip increases as a logarithmic function of time (provided it is correct to assume that friction is rate dependent).

The assumption that the rheology depends only on the slip rate might not be correct in reality, as the rheology could depend on other variables, such as cumulative slip since the last event, as well as on state variables that could coevolve with slip, such as temperature or pore pressure. In any case, rheological models of creeping faults have to satisfy the law expressed by Equation 8.

### 5. Dynamic Models of the Seismic Cycle

#### 5.1. Earthquakes and Friction Laws

The recognition that earthquakes probably represent frictional slip instabilities (Brace & Byerlee 1966) has spurred an interest in determining the frictional properties of a number of common fault-zone lithologies (Blampied et al. 1995, Dieterich 1979a,b; Ikari et al. 2009; Reinen et al. 1991). The stability of frictional sliding depends on whether frictional resistance increases or drops during slip and, if it drops, on the amplitude of elastic unloading during slip. In principle, earthquakes can nucleate only on fault patches where the frictional resistance drops during sliding. Such patches can creep, however, if the driving stresses drop even more rapidly during sliding and are said to be conditionally stable (Scholz 1998). Unstable slip cannot develop spontaneously on patches where the frictional resistance increases during sliding. Laboratory experiments have shown that frictional sliding is primarily a rate-dependent process in a steady-state regime (constant stress and sliding velocity) and that a state variable needs to be introduced in order to allow for
RATE-AND-STATE FRICTION

Laboratory experiments on frictional sliding have revealed that for most materials (for a review, see Marone 1998), the following conditions are true:

■ The resistance to sliding depends on the sliding rate at steady state, with a logarithmic dependency of the coefficient of friction on the slip rate.
■ The resistance to sliding increases to a transient peak value when an imposed slip rate is suddenly changed, and the peak value is a logarithmic function of the slip rate.
■ The friction coefficient increases approximately linearly with the logarithm of time in hold-and-slip experiments.

Laboratory measurements at low sliding rates, on the order of microns per second, can be reproduced relatively well with a rate-and-state formalism. Various laws have been proposed (Dieterich 1979a, Marone 1998, Ruina 1983). One common such law is the aging law, which writes

\[
\mu = \mu_s + a \ln \frac{V}{V_s} + b \ln \frac{\theta}{\theta_s} + \frac{d\theta}{dt} = 1 - \frac{V \theta}{D_c}.
\]

At steady state, the law is purely rate dependent:

\[
\mu_{ss} = \mu_s + (a - b) \ln \frac{V}{V_s}.
\]

The stability analysis of the aging law for a single-degree-of-freedom system, a spring-and-slider system, shows that slip can be stable only if \(a - b > 0\) and that unstable slip (stick-slip) requires \(a - b < 0\) (Scholz 1990). Unstable slip occurs if \(a - b\) is smaller than a critical negative value, defining an intermediate domain of conditional stability.

In the case of a crack with size \(L\) embedded in an elastic medium with shear modulus \(G\), the condition for unstable slip is

\[
b - a > \frac{\lambda GD_c}{L \sigma'_{\sigma}}.
\]

where \(\lambda\) is on the order of unity and \(\sigma'_{\sigma}\) is the effective normal stress (\(\sigma'_{\sigma} = \sigma_n - P\), where \(P\) is pore pressure).

Note that, in the limit when the pore pressure becomes near lithostatic, the critical value becomes infinite. An implication is that high pore pressure should promote stable slip though reduction of the effective normal stress.

(a) the transient behavior observed in non-steady-state experiments and (b) healing in hold-and-stick experiments (see sidebar, Rate-and-State Friction).

The rate-and-state formalism derived from these experiments has allowed for the development of numerical models of the seismic cycle. Early studies demonstrated that the formalism can reproduce qualitatively various aspects of fault behavior (Dieterich 1987, 1994; Tse & Rice 1986). The formalism predicts, for example, that afterslip should increase as a logarithmic function of time (Marone et al. 1991); it also predicts that earthquakes initiate with a phase of nucleation, as a sliding crack needs to reach a critical size before sliding becomes unstable (Ampuero & Rubin 2008, Dieterich 1987, Rubin & Ampuero 2005). This prediction is consistent with the observation that the onset of seismic slip on natural faults must be a time-dependent process to explain the negligible correlation between earthquakes and solid Earth tides (e.g., Beeler & Lockner 2003). This feature alone could also explain the Omori law of aftershock decay (Dieterich 1994).
A key element in models of the seismic cycle based on rate-and-state friction is the spatial
distribution of the rate dependency of friction \((a - b)\). Early two-dimensional models assumed
simple variation with depth, with \(a - b\) being homogeneously negative in the seismogenic depth
range and positive elsewhere. If \(a - b\) is assumed positive below depth \(z_d\) and negative above that,
the model can predict long interseismic periods of locking of the updip portion of the interface
and occasional rupture of the shallow portion, with transient afterslip within the transition zone
(e.g., Lapusta & Rice 2003, Lapusta et al. 2000, Tse & Rice 1986). Such models predict a bimodal
distribution of seismicity, with a population of small earthquakes confined to the transition zone
where stress builds up in the interseismic period as well as larger events that generally rupture the
whole rate-weakening patch. The modeled behavior is somewhat similar to the conceptual models
depicted in Figures 1 and 2. Depending on the assumed parameters, these models can predict more
or less overlap between the seismic ruptures and the zone of interseismic and postseismic creep.
The expectation is that the interseismic coupling and seismic coupling should be approximately 1
within the seismogenic depth range.

The Himalayan example would display behavior closest to that predicted by such models.
However, we have seen that the other examples show quite different behavior. Figure 11 illustrates
a revised conceptual model that would be more consistent with these observations. The key element
is that friction on the fault is now spatially heterogeneous, with interlacing rate-weakening and

---

**Figure 11**

Conceptual model of how the pattern of friction on a fault influences seismic ruptures and aseismic creep. The fault is interlaced with
rate-weakening and rate-strengthening areas. This pattern will determine the patterns of interseismic coupling, seismic ruptures, and
afterslip.
rate-strengthening patches. The fault is mostly seismic in areas where the rate-weakening patches are closely packed and mostly aseismic where they are sparse. The rate-weakening patches form asperities where stress builds up as the surrounding rate-strengthening areas creep. Assuming stress variations represent a modest fraction of the ambient stress level, at sufficient distances from these asperities creep will proceed at a uniform rate dictated by the ambient stress. As stress builds up on the asperities, they will cast a local stress shadow so that clustered rate-weakening patches will tend to form a zone with nearly uniform full locking ($\chi_i = 1$). Earthquakes are expected to initiate at the edge of those locked patches. The extent of the rupture will essentially depend on the stress distribution at the onset of the rupture and on the energy dissipated as seismic slip propagates into rate-strengthening areas. Afterslip would then result from aseismic creep of the rate-strengthening areas. The model allows the interseismic and seismic coupling coefficient to be <1 within the seismogenic depth range and produces a diversity of rupture scenarios. The expectation is that adjacent rate-weakening patches can sometimes fail together and sometimes fail independently, depending on the details of the stress distribution at the time of rupture. If heterogeneities of small size are allowed, this model can explain other behaviors, including SSEs (due to nucleation on the patches that is smaller than the earthquake nucleation size) (Liu & Rice 2007, Rubin 2008) and tsunami earthquakes (Bilek & Lay 2002).

5.2. Some Examples of Dynamic Simulations

The slip history predicted in the case of a single rate-weakening patch within a creeping fault assumed to obey rate-and-state friction (Ader et al. 2014) is shown in Figure 12. This model predicts a bimodal distribution of magnitudes. The smaller events, which are localized in the zone of interseismic stress and build up at the edges of the rate-weakening patch, follow the Gutenberg–Richter distribution, showing that the variety of rupture sizes revealed by the Gutenberg–Richter law can arise from the dynamics predicted by such a model without the need for geometric complexities. The larger events rupture the whole patch and show less diversity, occurring quasi-periodically and at a frequency that is inconsistent with extrapolation of the Gutenberg–Richter law defined by the smaller events. In an attempt at characterizing the influence of spatial variations of friction, Kaneko et al. (2010) analyzed a similar numerical experiment but with two rate-weakening patches separated by a rate-strengthening patch. If the intervening patch is strongly rate-strengthening or if its width, $W$, is sufficiently large, it acts as a persistent barrier to seismic rupture propagation. The product $W(a - b)\sigma_n$ (where $a$ and $b$ refer to the frictional properties of the rate-strengthening patch and $\sigma_n$ is the effective normal stress) determines both interseismic coupling and the probability that a seismic rupture spans the two rate-weakening patches. If this factor is small, the two patches behave as a single rate-weakening patch. Dublanchet et al. (2013) carried out a similar analysis in which they considered rate-weakening circular patches with identical radius distributed at random on a two-dimensional fault. In the context of this model, the critical width of Kaneko et al. (2010) translates to a critical density of asperities, above which seismic ruptures occasionally rupture the whole cluster.

These numerical experiments show that the conceptual model shown in Figure 11 can explain many of the observations reviewed here. The model can explain the diversity of earthquake ruptures and the relative regularity and similarity of the largest events, the occurrence of both seismic and aseismic slip associated with a single event, the observation that interseismic and seismic coupling averaged over the seismogenic zone are <1, and the observation that seismic ruptures are generally confined to zones of high coupling. In principle, it should be possible to design numerical models tuned to reproduce observations from a particular fault system. For example, a cluster of rate-weakening patches embedded in a creeping zone significantly larger than their stress...
Figure 12
Simulation of fault slip in the case of a single rate-weakening ($a - b < 0$) small patch embedded in a rate-strengthening ($a - b > 0$) fault, based on Ader et al. (2014). The fault is infinite in one direction and is loaded at a constant slip velocity $V = 1$ cm/yr on both sides. See Ader et al. (2014) for details. (a) Isochrons of slip. If the fault is experiencing a seismic event, the slip on the fault is plotted every 0.02 s (dashed red lines); during the interseismic period, slip is plotted every 0.01 yr (solid blue lines). (b) Magnitude distribution of events produced by the fault. $N$ is the number of events at a given magnitude. (c) Magnitude-frequency distribution of seismic events (the ordinate is the logarithm of the number of events with magnitude above the value indicated on the abscissa).
shadow could produce maximum events that would repeat quite regularly as well as smaller, more irregular events. Such a model could be tuned to reproduce examples of repeating earthquakes such as the Kamaishi sequence (e.g., Uchida et al. 2012) or the sequence of $M_c$ 6.0 earthquakes on the Parkfield segment of the San Andreas Fault. Wang et al. (2012b) carried out an inversion of the geodetic data covering the pre-, co-, and postseismic periods at Parkfield and found that a self-consistent model could be obtained in which coseismic strain released interseismic strain and drove postseismic strain. Their solution indicates a well-defined locked patch that partially ruptured in the $M_c$ 6.0 earthquake of 2004. Barbot et al. (2012) produced a friction model tuned to reproduce the pattern of interseismic strain, seismic events similar to the 2004 earthquake, and postseismic strain (Figure 13). They defined the rate-weakening patches on the basis of the observed pattern of interseismic coupling, background seismicity, and aftershocks. The model produced a realistic sequence of synthetic earthquakes with a diversity of rupture distributions and return periods of the larger events similar to observations. In my opinion, these simulations and other similar studies (e.g., Hori et al. 2004, Kato 2008) demonstrate the power of dynamic modeling not only to validate concepts but also to produce realistic synthetic catalogs. Such simulations can be useful to evaluate the full range of possible earthquake scenarios that a particular fault system could host. There are, however, important caveats: First, such models are highly nonunique, especially if only limited observations are available to constrain the model, and second, there is no proof that these models are based on the right physics.
5.3. Limitations

The ideal earthquake simulator would be capable of using realistic fault physics to dynamically model earthquake cycles on large-scale faults such as subduction zones and on complex systems of multiple faults. Several aspects of fault physics make this goal intractable with current computational resources. For example, techniques exist to simulate dynamic propagation of a single event on a fault with nonplanar geometry (e.g., Aochi et al. 2000, Oglesby & Day 2001, Oglesby & Mai 2012), but fully dynamic simulation of earthquake cycles on faults with three-dimensional geometries remains out of reach, necessitating simplifications (e.g., Duan & Oglesby 2005, Hori et al. 2004) that might bias modeling results (Thomas et al. 2014c).

Laboratory observations at seismic sliding rates (e.g., Di Toro et al. 2011) additionally suggest a stronger weakening mechanism during seismic sliding than predicted by the aging rate-and-state friction law outlined above. Intense weakening due to thermal pore pressurization, for example, may explain why the Tohoku-oki earthquake was able to rupture the shallow, presumably rate-strengthening portion of the Japan megathrust (Noda & Lapusta 2013). Laboratory experiments (e.g., Brantut et al. 2008) and field observations (Rowe et al. 2012a) indicate that fault-zone rocks can also undergo metamorphism during seismic sliding (e.g., Brantut et al. 2008, Rowe et al. 2012a), which could significantly affect the dynamics of seismic faulting in nature. Another limitation is the difficulty of calibrating models in the absence of observations covering a complete seismic cycle, as was the case in the example from the Nepal Himalaya introduced above. Nevertheless, although fully dynamic simulations are still far from capturing the full range of processes that might be at play in reality, they offer a promising avenue to develop realistic simulations of the seismic cycle that can be tuned to geodetic and seismological observations.

6. CONCLUSIONS AND SOME OUTSTANDING ISSUES

Recent advances in GPS geodesy and remote sensing make it possible to monitor strain buildup and release on faults. Assuming that geodetic strain is only elastic (or is corrected for strain due to known earthquakes), these measurements can be used to draw inferences regarding the possible frequency and magnitude of the most extreme events. The underlying principle is that seismic ruptures and postseismic processes release strain that matches that accumulated over the interseismic period.

The view that slip within the seismogenic zone is only seismic seems to apply to the MHT and may be a reasonable hypothesis for most continental faults. This view does not apply to subduction zones or to faults within the oceanic crust and in areas of hydrothermal activity. The mechanisms involved in aseismic fault slip and the factors determining the partitioning between seismic and aseismic slip in time and space remain elusive. Temperature, lithology, and fluids are certainly key factors, but our understanding of the physical mechanisms involved is insufficient to derive a general theory of fault creep.

The hypothesis that interseismic strain is only elastic is conservative with regard to seismic hazard assessment. It may in fact be an overly conservative hypothesis in slowly deforming regions where distributed anelastic deformation could take up a large fraction of the measured geodetic strain, although the occurrence of large earthquakes away from plate boundaries demonstrates that the effective viscosity within the plate is too small to prevent elastic strain buildup. The approach presented in this article is best adapted to study very active fault zones with rapid geodetic strain rates, where this issue can be neglected.

Dynamic models might help in the future to assess the range of earthquakes that a fault system could produce. However, calibrating such models in the absence of geodetic and seismological...
observations covering the whole seismic cycle is a major challenge. If the approach turns out to be practical, such models might provide guidance to help estimate the probable rupture area and slip distribution frequency of the most extreme events. If techniques were developed to probe not only stress buildup (as with geodesy) but also the spatial distribution of absolute stresses, we would be one small step away from being able to estimate a time-dependent probability based on dynamic modeling, opening the possibility for a weather forecasting approach to seismic hazard.

SUMMARY POINTS

1. The patterns of locking on megathrusts, and on some continental faults, are highly heterogeneous. Seismic ruptures are mostly confined within patches that remain locked in the interseismic period.

2. Geodetic measurements of interseismic strain can be used to place constraints on the frequency-magnitude relationship of the maximum-magnitude earthquake.

3. A large fraction of long-term slip within the seismogenic zone is aseismic on megathrusts, oceanic transform faults, and faults at oceanic spreading ridges. Creep is prevalent on faults with clay-rich gouge and faults in hydrothermal systems.

4. Postseismic slip increases as the logarithm of time. This observation suggests that steady-state friction on creeping faults depends on the logarithm of slip velocity, as observed in the laboratory for most materials.

5. The frequency-magnitude law determined from smaller earthquakes might be used to estimate the magnitude and frequency of the maximum-magnitude earthquake.

6. Aseismic afterslip drives aftershocks in general.

7. Dynamic models of the seismic cycle based on rate-and-state friction can reproduce many of these observations. If such models are validated and calibrated they might be used to assess the characteristics of the most extreme events for seismic hazard studies.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I am deeply indebted to the many colleagues, students, and postdoctoral fellows who worked with me on the topics covered in this article. I thank in particular my colleagues Kerry Sieh, who initiated and led the project on the Sumatra megathrust; Nadia Lapusta, who supervised the dynamic modeling studies; Hugo Perfettini, with whom I have had an enjoyable collaboration to investigate creeping faults; and Don Helmberger, whose seismological methods and insight have been critical. I also thank Shengji Wei, Andrew Kositsky, Ozgun Konca, Marion Thomas, Thomas Ader, Mohamed Chlieh, Anthony Sladen, Yoshi Kaneko, Yaru Hsu, Sylvain Barbot, Mark Simons, and Pablo Ampuero. Lisa Christiansen and Paul Avouac helped improve most of the graphics. I thank Chris Rollins and Jocelyn Rice for their help in clarifying and editing the text and Doug Beckner for editing the graphics.
LITERATURE CITED


Avouac JP. 2011. The lessons of Tohoku-Oki. Nature 475:300–1


Hyndman RD, Yamano M, Oleskevich DA. 1997. The seismogenic zone of subduction thrust faults. Island Arc 6:244–60
Irwin WP, Barnes I. 1975. Effect of geologic structure and metamorphic fluids on seismic behavior of the San Andreas fault system in central and northern California. Geology 3:713–16


Contents

A Conversation with James J. Morgan  
James J. Morgan and Dianne K. Newman ...................................................... 1

Global Monsoon Dynamics and Climate Change  
An Zhisheng, Wu Guoxiong, Li Jianping, Sun Youbin, Liu Yimin, Zhou Weijian,  
Cai Yanjun, Duan Anmin, Li Li, Mao Jianguo, Cheng Hai, Shi Zhengguo,  
Tan Liangcheng, Yan Hong, Ao Hong, Chang Hong, and Feng Juan .................. 29

Conservation Paleobiology: Leveraging Knowledge of the Past  
to Inform Conservation and Restoration  
Gregory P. Dietl, Susan M. Kidwell, Mark Brenner, David A. Burney,  
Karl W. Flessa, Stephen T. Jackson, and Paul L. Koch .................................. 79

Jadeitites and Plate Tectonics  
George E. Harlow, Tatsuki Tsujimori, and Sorena S. Sorensen .......................... 105

Macroevolutionary History of the Planktic Foraminifera  
Andrew J. Fraass, D. Clay Kelly, and Shanan E. Peters ................................. 139

Continental Lower Crust  
Bradley R. Hacker, Peter B. Kelemen, and Mark D. Behn ............................... 167

Oceanic Forcing of Ice-Sheet Retreat: West Antarctica and More  
Richard B. Alley, Sridhar Anandakrishnan, Knut Christianson, Hau J. Horgan,  
Atsu Muto, Byron R. Parizek, David Pollard, and Ryan T. Walker .................... 207

From Geodetic Imaging of Seismic and Aseismic Fault Slip  
to Dynamic Modeling of the Seismic Cycle  
Jean-Philippe Avouac  ..................................................................................... 233

The Pyrogenic Carbon Cycle  
Michael I. Bird, Jonathan G. Wynn, Gustavo Saiz, Christopher M. Wurster,  
and Anna McBeath ..................................................................................... 273

The Architecture, Chemistry, and Evolution of Continental  
Magmatic Arcs  
Mihai N. Ducea, Jason B. Saleeby, and George Bergantz ............................... 299

Paleosols as Indicators of Paleoenvironment and Paleoclimate  
Neil J. Tabor and Timothy S. Myers ............................................................. 333
Role of Arc Processes in the Formation of Continental Crust
 Oliver Jagoutz and Peter B. Kelemen ........................................... 363

Environment and Climate of Early Human Evolution
 Naomi E. Levin ............................................................................. 405

Magma Fragmentation
 Helge M. Gonnermann .................................................................. 431

Atmospheric Escape from Solar System Terrestrial Planets and Exoplanets
 Feng Tian ....................................................................................... 459

A Tale of Amalgamation of Three Permo-Triassic Collage Systems in Central Asia: Oroclines, Sutures, and Terminal Accretion
 Wenjiao Xiao, Brian F. Windley, Shu Sun, Jiliang Li, Baobin Huang, Chunming Han, Chao Yuan, Min Sun, and Hanlin Chen ......................... 477

Atmospheric Dynamics of Hot Exoplanets
 Kevin Heng and Adam P. Showman ............................................ 509

Transient Creep and Strain Energy Dissipation: An Experimental Perspective
 Ulrich Faul and Ian Jackson ....................................................... 541

Rapid Plate Motion Variations Through Geological Time: Observations Serving Geodynamic Interpretation
 Giampiero Iaffaldano and Hans-Peter Bunge .................................. 571

Rethinking the Ancient Sulfur Cycle
 David A. Fike, Alexander S. Bradley, and Catherine V. Rose ................ 593

Indexes
 Cumulative Index of Contributing Authors, Volumes 34–43 .................. 623
 Cumulative Index of Article Titles, Volumes 34–43 ............................. 628

Errata
 An online log of corrections to Annual Review of Earth and Planetary Sciences articles may be found at http://www.annualreviews.org/errata/earth

Contents ix