Understanding the process of faulting: selected challenges and opportunities at the edge of the 21st century

R.A. Schultz

Geomechanics–Rock Fracture Group, Department of Geological Sciences, Mackay School of Mines/172, University of Nevada, Reno, NV 89557, USA

Received 18 March 1998; accepted 12 January 1999

Abstract

Displacement distributions along fault surfaces are a record of the processes of fault nucleation, slip, linkage, and propagation. Because several disparate processes can collectively influence the fault displacements, the relative contributions of these processes can be challenging, but are therefore necessary, to decipher. Continued advances in the mechanics of discontinuous slip surfaces in rock masses, when combined with appropriate field studies, should determine the roles of cohesive end zones, fault interaction, and propagation direction in shaping the displacement distributions. Studies of faults on other planetary surfaces provide a window into the role of a broad range of environmental conditions that can influence the faulting process. More inroads need to be made into traditional strain-based classes in structural geology so that mechanically sound concepts of fault analysis can become better utilized in the curriculum and by the non-specialist geological community.

1. Introduction

The past decade has witnessed explosive growth in our understanding of faults. Focused studies on the process of faulting have largely supplanted a reliance on various classification schemes that have been in use during this century (e.g. Reid et al., 1913; Billings, 1972; cf. Johnson, 1995) and have provided a useful framework for integrating faults into regional-scale contexts (e.g. Atwater, 1970; Nur et al., 1986; Jackson et al., 1988; Angelier, 1994). The newly emerging field of Geologic Fracture Mechanics (GFM), an interdisciplinary field that combines approaches from engineering, materials science, and geology, provides a basis for understanding the detailed behavior of faults (and other types of fractures such as joints: Pollard and Aydin, 1988) as entities unto themselves (e.g. Aydin and Johnson, 1983; Pollard and Segall, 1987; Cowie and Scholz, 1992b; Engelder et al., 1993).

The simplest formulation of fractures as slits in an otherwise homogeneous, isotropic, linearly elastic material (Linear Elastic Fracture Mechanics, LEFM) has provided a first-order understanding of how geological fractures accommodate displacement, concentrate stresses, interact, and propagate (see Pollard and Segall, 1987 and Engelder et al., 1993 for clear treatments). For example, the controls on discontinuous joint and fault geometry, and their échelon configurations, are well understood through calculations of the elastic stress interactions between them (e.g. Segall and Pollard, 1980; Pollard and Aydin, 1988; Martel and Pollard, 1989; Aydin and Schultz, 1990; ten Brink et al., 1996; Willemse et al., 1996; Crider and Pollard, 1998). However, field work during the past decade has demonstrated significant departures from this ideal behavior, especially in revealing tapered, non-elastic near-tip displacement profiles along dikes (Rubin, 1993; Ward, 1993) and faults (Cowie and Scholz, 1992b; Schultz and Moore, 1996; Cartwright and Mansfield, 1998; Moore and Schultz, 1999). Parallel advances in modeling capability showed the importance of three-dimensional fracture shape and segment geometry (Willemse, 1997; Crider and Pollard, 1998).
position-, time-, and slip-dependent friction along a fault (Cowie and Scholz, 1992b; Willemse et al., 1996; Cooke, 1997; Martel, 1997), and inhomogeneities such as variations in lithology or rheology (Nicol et al., 1996; Gudmundsson, 1998) or remote stress state (Bürgmann et al., 1994) on the fault displacements. The elastic–plastic and three-dimensional models of geological fractures provide a significant and necessary improvement over simpler LEFM-based approaches for relating fracture-scale processes to crustal-scale tectonics.

The coming decade will continue to see progress made in the quantitative understanding of faults and the processes associated with their nucleation, stability, accumulation of displacements, interaction, and propagation. With this background, a more refined and mechanically based understanding of brittle deformation of the Earth's crust may be motivated. In particular a closer integration of continuum physics, appropriate to problems of continental scale (e.g. Heimpel and Olson, 1996), with the mechanics of three-dimensional networks of discontinuous faults, may lead to a tighter integration between structural geology and geodynamics at the appropriate relative scale (Schultz, 1996).

In this note, I touch on some critical problems in faulting that presently challenge our understanding of this process. Although many areas could be examined at length, I choose to highlight three classes of problems: competing controls on fault displacements; extension of fault mechanics studies to other planetary surfaces; and diffusion of fault mechanics concepts to the broader geological community. Each of these areas presents distinct opportunities to structural geologists. Other topics not developed here for reasons of brevity or expertise are treated elsewhere in this volume.

2. Progress in fault mechanics

The investigation of faults as the sites of repeated frictional slip events has blossomed and matured in recent years, with an extensive literature in many areas. For example, experimental and theoretical investigation of the frictional properties of fault surfaces (e.g. Dieterich, 1978; Tse and Rice, 1986; Scholz, 1990; Segall, 1991; Marone, 1998) has shown how slip velocity, duration of stationary contact, rock-mass stiffness, and temperature (among other factors) can influence the reduction (slip weakening) or recovery (fault healing) of frictional strength and its stability (seismic slip or stable creep). This work is necessary, in part, to relate field observations of faults—such as their near-tip displacement profiles—to more comprehensive theoretical models of geological fractures, such as those that incorporate an explicit zone of plastic deformation near fracture tips ('cohesive zones,' discussed later).

Mechanical analyses of faults as slipping cracks in elastic media, in two dimensions, provide a baseline for understanding the kinematics of fault interaction and linkage as well as the discontinuum mechanics of crustal deformation (e.g. Chinnery, 1961; Segall and Pollard, 1980; Aydin and Nur, 1982; Aydin and Scholz, 1990; King and Ellis, 1990; Scholz and Aydin, 1990; Cowie and Scholz, 1992b; Harris and Day, 1993; Ohlmacher and Aydin, 1997). More recently, three-dimensional modeling has demonstrated the importance of non-plane geometries on fault kinematics and mechanics (Katzman et al., 1995; ten Brink et al., 1996; Willemse et al., 1996; Willemse, 1997; Crider and Pollard, 1998; Martel and Boger, 1998). Field and numerical investigations of the kinematic consequences of fault linkage (e.g. Peacock, 1991; Dawers et al., 1993; Davison, 1994; Peacock and Sanderson, 1991, 1994; Trudgill and Cartwright, 1994; Faulds and Varga, 1998), of fault slip in neotectonics and seismic hazard (e.g. Wells and Coppersmith, 1994; Wesnousky, 1988; Sibson, 1989; Scholz, 1990; Main, 1996), and of fault-population statistics and mechanics (e.g. Cowie et al., 1993, 1995; Wojtal, 1994; Cartwright et al., 1995; Cladouhos and Marrett, 1996; Marrett, 1996) underscore the common fundamental physics. One key element common to all these areas is the displacement distribution along the fault, particularly near the tip.

Gradients in slip along a fault's trace (map) length have become well documented during the 20th century (discussed for example by Segall and Pollard, 1980; Muraoka and Kamata, 1983; Barnett et al., 1987; Walsh and Watterson, 1989; Scholz, 1990). More recent studies have focused on the sequence of fault slip, interaction, propagation, and linkage inferred from these distributions (e.g. Peacock and Sanderson, 1991; Dawers et al., 1993; Davison, 1994; Cartwright et al., 1995; Willemse, 1997; Cartwright and Mansfield, 1998; Cowie and Shipton, 1998). As a result, a basic understanding of the mechanics of fault segmentation and linkage has been established for a variety of simple fault configurations. One challenge in fault analysis is to determine and quantify the responsible elements.

The displacement distribution along a fault is influenced by several factors, including:

1. fault (trace) length in map or cross-sectional view (e.g. Segall and Pollard, 1980; Muraoka and Kamata, 1983; Cowie and Scholz, 1992a);
2. fault aspect ratio (trace-length/down-dip height; Willemse et al., 1996; Nicol et al., 1996);
3. fault shape (rectangular vs elliptical; Sih, 1975; Scholz, 1982; Barnett et al., 1987; Willemse et al.,
4. frictional and constitutive properties of the fault (Lin and Parmentier, 1988; Aydin and Schultz, 1990);  
5. near-tip processes (Palmer and Rice, 1973; Li, 1987; Cowie and Scholz, 1992b);  
6. mechanical interaction with other faults (e.g. Segall and Pollard, 1980);  
7. fault-segment linkage (Davison, 1994);  
8. configuration of far-field stresses (Bürgmann et al., 1994);  
9. elastic properties and variations in lithology displaced by the fault (Scholz et al., 1993; Bürgmann et al., 1994);  
10. proximity to the free surface and other boundaries (e.g. Bruhn and Schultz, 1996; Crider and Pollard, 1998);  
11. inter-fault plate deformation (King and Ellis, 1990; Moore and Schultz, 1999); and  
12. time-dependent faulted rheologies (Tse and Rice, 1986; Freed and Lin, 1998).

Because the displacement distributions that we observe and measure on faults, as well as the related displacement–length scaling relations, depend on one or more of the above, a systematic investigation of these (and others), for each field example, appears necessary for an understanding of fault displacement profiles to be achieved.

For example, uniform lithologic and remote stress conditions can lead to symmetric slip distributions along faults (Bürgmann et al., 1994); however, the location of the maximum displacement along the fault does not appear to be a reliable indicator of the point of fault ‘nucleation’ given the results of the previous studies. Non-uniform conditions, such as stress gradients acting on the fault, variations in frictional properties, polylithologic rock masses, and interaction with the Earth’s surface or other faults, can each lead to asymmetric displacement distributions, and steeper displacement gradients, along faults (e.g. Segall and Pollard, 1980; Bürgmann et al., 1994; Bruhn and Schultz, 1996; Willemse et al., 1996).

The near-tip region of a fault is particularly interesting as it is the nexus of processes critical to strain accumulation in the crust. Here, the accumulation of displacements associated with frictional sliding and spread of rupture patches along the fault (e.g. Martel and Pollard, 1989), and formation of process zones (e.g. King and Yielding, 1984; Li, 1987) and secondary structures (e.g. Cooke, 1997; Martel, 1997; Ohlmacher and Aydin, 1997) that modulate the elevated stress levels beyond the tip, can often be observed in the field.

Recent studies have documented tapering or linearly varying near-tip displacement profiles for faults over a range of scales (e.g. Cowie and Scholz, 1992b; Schultz and Moore, 1996; Cartwright and Mansfield, 1998; Cowie and Shipton, 1998). These profile shapes are interesting and important in that they demonstrate inelastic (plastic) strains at the tip. Current work demonstrates that fault interaction (‘soft linkage’) can promote tapering displacement profiles in the step over (‘relay-ramp’) regions between échelon faults, especially for surface-breaking normal faults (e.g. Peacock and Sanderson, 1991; Davison, 1994; Trudgill and Cartwright, 1994; Willemsen et al., 1996; Willemsen, 1997; Crider and Pollard, 1998). In contrast, however, tapering profiles of the fault tips themselves appear to be required by the distribution and patterns of secondary structures such as splay cracks (Cowie, 1997; Martel, 1997), relatively low rock strengths (Cowie and Shipton, 1992b; Schultz and Moore, 1996), and displacement–length scaling relations (Cowie and Scholz, 1992a; Clark and Cox, 1996; Cowie and Shipton, 1998); comparable tapered profiles are identified, and interpreted mechanically in a similar fashion, for dilatant cracks and dikes (e.g. Pollard et al., 1982; Nicholson and Pollard, 1985; Rubin, 1993).

Both fault interaction and the mechanics of fault terminations likely influence the step over kinematics and relay-ramp configurations discussed earlier, making field and numerical studies of fault step overs of prime importance. Additional measurements of near-tip profiles for faults are needed to extend the database to a wider range of natural examples. Explicit modeling of échelon faults having inelastic (i.e. elastic–plastic) properties would also aid in separating these effects.

The stress state (or strain energy density) in the near-tip region governs the formation of secondary structures, leading to ‘hard linkage’ between fault segments (e.g. Martel and Pollard, 1989; Davison, 1994; Trudgill and Cartwright, 1994; Crider and Pollard, 1998), or in-plane propagation of the faulted surface (likely in mode-III: e.g. Cowie and Shipton, 1998; cf. Lin and Parmentier, 1988; Du and Aydin, 1993). A careful distinction is usually drawn between the propagation (or spread) of rupture patches along a fracture surface (e.g. Martel and Pollard, 1989) and the propagation of the surface itself. Although controversial, the search for localized ‘damage zones’ or ‘process zone wakes’ thought to be induced in certain rock types surrounding a fault by passage of a rupture patch (e.g. Scholz et al., 1993; Cowie and Shipton, 1998; Vermilye and Scholz, 1998), along with critical assessments of near-tip kinematics and fracturing, are fruitful areas for further advances in fault propagation studies. A significant challenge involves establishing a meaningful correspondence between field observations of particular fault tip regions and more generic mechanical representations.
of this region obtained from elastic–plastic (e.g. cohesive-zone) formulations.

3. Opportunities in planetary structural geology

Populations of faults on planetary surfaces (beyond the Earth) provide an additional and unusually informative window into the process of faulting. Each planetary body developed under specific conditions that may differ substantially from those characteristic of the Earth. The lack of significant atmosphere on Mercury, the Moon, and most icy satellites, combined with exceedingly slow erosion rates (associated with an absence of fluvial, pluvial, eolian, and hydrological processes), permits preservation of unusually clear fault morphologies. Given a lack of crustal recycling and Earth-like plate tectonics on most planetary bodies, a visible record of single or superposed faulting episodes may be preserved, revealing details of the development of the fault populations—including fault-segment linkage—over several orders of magnitude of length. Differences in planetary surface gravities, and surface temperatures, imply that the faulting developed under a broad range of confining pressures, rock-mass stiffnesses, and rates. As a result, planetary surfaces provide unique natural laboratories for studying the process of faulting under a wider range of environmental conditions (gravity, pore-water pressure, temperature, tectonic regime) than is possible by using terrestrial fault sets alone.

Faults have been documented on nearly every geological surface in the solar system (e.g. Greeley, 1994) and a vast literature exists on the subject of planetary structural geology. Normal fault and graben systems are probably the most common (Figs. 1a and b), accommodating both localized and regional extension on Mercury, Venus (e.g. Hanson and Willis, 1996), the Moon (e.g. Golombek, 1979), Mars, Europa, Ganymede, and several smaller icy satellites of the outer planets including Tethys, Dione, and Miranda (e.g. Pappalardo et al., 1997). Thrust faults have been identified on several planets and satellites including Mercury (Figs. 1d and e), Venus and Mars (e.g. Suppe and Connors, 1992; Watters, 1993; Watters et al., 1998). Structures known as ‘wrinkle ridges’ (e.g. McGill, 1993), thought to accommodate contractional strain, define networks (e.g. Watters, 1992) and populations on Mercury, Venus, the Moon, Mars, and perhaps some icy satellites. Strike-slip faults have been identified on Mars (Schultz, 1989; Fig. 1c) and on some icy satellites (e.g. Schenk and McKinnon, 1989) although large lateral displacements such as those found systematically at terrestrial transform plate boundaries may not be as common on other planets as on Earth (see Koenig and Aydin, 1998, for a rare

![Fig. 1. Examples of bounded faults on planetary surfaces. (a) Échelon grabens from the Alba Fossae array of Mars, located near 35°N, 120°; Viking orbiter image 252S22, north toward the top. Note obvious right step over between grabens and tapering of throw along bounding faults (after Schultz, 1997). U, up; D, down. (b) Échelon grabens of the Devils Lane system, Canyonlands National Park, Utah. Two nearly orthogonal joint sets predate the normal faulting. (c) Échelon strike-slip faults on Mars (after Schultz, 1989). Although no offset markers along fault segments are apparent, the left-lateral slip (arrows) can be deduced by inspection from the right-stepping segment geometry along with the contractional sense of strain within the step overs. Note uplifted secondary ridges at segment terminations and composite (multiple fault strand) nature of contractional step over. Viking orbiter image 610A27; north toward upper left. (d) Bounded arcuate thrust fault on Mercury showing contractional strain across scarp; total scarp length 500 km (Carr, 1984, pp. 48), scarp height ≈2 km. Mariner 10 image 528884. (e) Part of Discovery Rupes thrust fault on Mercury showing contractional strain across scarp; total scarp length 500 km (Carr, 1984, pp. 48), scarp height ≈2 km. Mariner 10 image 528884.]
example on Venus and Pappalardo et al., 1998, for examples on Ganymede). At present the resolution of orbital spacecraft imaging systems is insufficient to resolve individual dilatant cracks (joints), although lineaments and certain erosional landforms on Mars and elsewhere have been interpreted as being joint controlled (Tanaka and Golombek, 1989).

Populations of faults are well developed on other planets and satellites although their recognition and analysis have lagged behind that of terrestrial fault populations. For example, only recently have structural offsets (displacements) along planetary faults been related systematically to the fault lengths (Schultz, 1995, 1997; Schultz and Fori, 1996; Watters et al., 1998). This work demonstrates that the (maximum) displacement–length scaling relationships obtained for terrestrial faults (e.g. Cowie and Scholz, 1992a; Clark and Cox, 1996; Marrett, 1996) apply equally well to the examples of planetary faults evaluated (Fig. 1), despite differences in environmental parameters such as gravity, noted earlier. Similarly, the length–frequency distribution of planetary faults, population exponents (Cladouhos and Marrett, 1996), and methods for estimating population strain (e.g. Marrett, 1996) developed for Earth-based datasets should apply, providing new tools for quantifying planetary brittle strain fields and relating them to geodynamic models.

One first-order problem in planetary geomechanics is an evaluation of the stability of frictional sliding along faults on other planetary bodies. For cases in which the frictional yield criterion is met (i.e. Byerlee’s rule on the brittle strength envelope; Kohlstedt et al., 1995), displacements can build up on surface-breaking faults either seismically or aseismically. The stability conditions associated with seismic/aseismic slip are complex (Marone, 1998), with contributions from fault-zone structure and rheology (e.g. gouge), velocity- (and time-) dependent friction, temperature, strength recovery (healing), and stiffness of the loading system (i.e. the surrounding rock mass). A simple three-layer system (e.g. Scholz, 1990, pp. 129; Marone, 1998) provides a context for the vertical distribution of planetary seismicity.

The maximum depth of seismic (‘unstable’) faulting may be approximated by the thermal stability transition (e.g. Tse and Rice, 1986; Scholz, 1990; Marone, 1998). Frictional sliding along faults can be stable (velocity strengthening) below this depth for temperatures above ~300°C (for felsic rocks, or ~250°C for mafic rocks; Stesky et al., 1974). Slip at cooler temperatures, above this depth, is expected to be velocity weakening, unstable (stick-slip), and therefore seismogenic (Scholz, 1990). The transition between the seismic regime and the shallower stable regime is called the ‘upper stability transition’ (Marone and Scholz, 1988). In this near-surface region, the competing effects of gouge evolution and fault slip may control the depth distribution of seismicity (Marone, 1998). For example, faults having slower slip rates [relative to the rates of healing (lithification and consolidation of gouge)] may be associated with velocity strengthening behavior and stable, creep-like frictional sliding in this regime, which may extend to depths of 3–5 km in the Earth (e.g. Scholz, 1990; Marone, 1998).

Depending on specific values of geothermal gradient (e.g. 10 < \( \delta T/\delta z < 0.8 \) km\(^{-1} \)), the lower, thermal stability transition should occur on planets and satellites at depths of less than perhaps 10–30 km. The depth of seismic slip in planetary lithospheres should range between this value and that associated with the depth of the upper stability transition. Dilatancy of unconsolidated fault gouge (and velocity strengthening behavior) in the uppermost regime depends in part on confining pressure, implying that planetary bodies having lower surface gravities, such as the Moon, Ganymede and other icy satellites, Mercury, or Mars, may permit stable frictional sliding to greater depths than larger (or denser) planetary bodies, such as Venus or Earth. These experimental studies imply a formidable ambiguity in ascertaining the seismic potential of surface-breaking planetary faults from imaging data alone. Future work that combines high-resolution images of planetary fault zones, appropriate laboratory studies of frictional stability and, perhaps, in-situ seismic experiments, may lead to testable predictions of the seismic potential of planetary fault populations once this problem is resolved.

4. Spreading the word—refining the curriculum in geologic fracture mechanics

Results from rock mechanics and the mechanical analysis of faults may deserve a fuller exposure in college and university level structural geology courses. Because many undergraduate students in geology may be either unprepared for the level of mathematical sophistication expected for the material or remain unconvinced of its usefulness, the ‘stress side’ of structure and tectonics remains opaque to many geologists, professors, or students who may be more comfortable with strain. Yet it is clear that an understanding of mechanics principles (e.g. Johnson, 1970) should lead to a more comprehensive understanding of the associated strains and rock-mass deformation. Results from neotectonics and seismotectonics (e.g. Scholz, 1990) should also be included routinely in required undergraduate courses given their demonstrable relevance to faulting. Widespread dissemination of planetary image and topographic data on the World Wide Web (e.g. http://photojournal.jpl.nasa.gov) makes planetary fault
populations more available to the structural geological community and to the undergraduate curriculum, although the accessibility of planetary data can be impeded by the rather obscure nomenclature and special problems inherent to planetary studies. Tighter collaboration and communication can only be a benefit to all populations of structural geologists.

A paradigm shift perhaps similar to that identified in seismotectonics (Scholz, 1992) appears timely for classroom-based studies of faulting, with a substantial move away from description and classification to a more mechanically based approach (Johnson, 1970; Means, 1976; Turcotte and Schubert, 1982; Schultz, 1998). Although many colleges and universities in the US (and elsewhere) have developed strong programs in mechanics-based structural geology, expansion of this approach to the majority of undergraduate departments should prove to be a challenging yet worthwhile long-term strategy. It may take some effort for professors and lab instructors to supplement rigid-block and card-deck models of shearing that rely on infinite, unbounded structures, for example, with discontinuum mechanics appropriate to bounded faults (e.g. Mandl, 1987; Segall and Pollard, 1980; Davison, 1994), but the attempt should perhaps be made. As a concrete example, no longer should a student answer with just a number (e.g. 4 m) when asked how much slip there is on a particular fault, but instead should specify either the maximum displacement or the value and relative position along the fault (e.g. center or near the fault termination), while demonstrating an understanding of the physics behind the variation. Such a paradigm shift, for opening-mode cracks, was advocated more than a decade ago by Pollard and Aydin (1988), who drew contrasts between continuum-based (Mohr circle) and statistical or pattern recognition-based approaches, and stress-based fracture mechanics approaches, to studying joint sets. A similar shift and re-evaluation of our undergraduate curricula should occur today to encompass a broader working knowledge of both crack and fault processes.

5. Concluding remarks

The discontinuous geometry of faults in three-dimensions exerts fundamental control on the kinematics and mechanics of fault sets on planetary surfaces as well as the geodynamics of brittle crustal deformation. Adequately representing the discontinuous geometries of faults is a critical element of contemporary structural analysis. Although considerable progress has been made by the use of LEFM approaches, further progress will be enhanced by an increased reliance on more comprehensive formulations of fault displacement accumulation and distribution that explicitly include elastic–plastic and time-dependent character. A close integration of field work, laboratory experiments, and theory will promote progress in this difficult and multidisciplinary field.

While several of the factors that control fault nucleation, interaction, propagation, and linkage are appreciated from a field or kinematic perspective, careful analytical or numerical simulations of these processes, guided by geological observations, provide an informative avenue for progress in the future. An explicit treatment of three-dimensional fault geometry, interactions with stratigraphy, and brittle strain fields in three dimensions should not only comprise a substantial research effort but should also become more familiar to undergraduate students in geology. This innovative approach would be facilitated by broad and rigorous training of students in field work, experimental methods, and appropriate theory instead of a more seductive focus on specialization.

Acknowledgements

Reviews of the manuscript by Bruno Vendeville, David Ferrill, Jim Evans, and, especially, the anonymous referee, have sharpened the ideas and focus of this paper. I thank Daniel Mége, Catherine Homberg, and Chris Marone in particular for helpful discussions. Grants from NASA’s Planetary Geology and Geophysics Program, the U.S. Department of Energy, Chevron Petroleum Technology Company, and NSF that have supported or facilitated several aspects of the work presented here are gratefully acknowledged. This paper was completed while the author enjoyed sabbatical-year positions at Woods Hole Oceanographic Institution, Massachusetts, and at the Université Pierre et Marie Curie, Paris VI, France.

References


Marti, S.J., Boger, W.A., 1998. Geometry and mechanics of second-


