Modeling shear zones in geological and planetary sciences: solid- and fluid-thermal–mechanical approaches

K. Regenauer-Lieb a,* D.A. Yuen b,1

a Institute of Geophysics, ETH Zürich, Hönggerberg HPP015-Gebäude, CH-8093 Zürich, Switzerland
b Department of Geology and Geophysics, Supercomputer Institute, University of Minnesota, Minneapolis, USA

Received 15 April 2002; accepted 4 March 2003

Abstract

Shear zones are the most ubiquitous features observed in planetary surfaces. They appear as a jagged network of faults at the observable brittle surface of planets and, in geological exposures of deeper rocks, they turn into smoothly braided networks of localized shear displacement leaving centimeter wide bands of “mylonitized”, reduced grain sizes behind. The overall size of the entire shear network rarely exceeds kilometer scale at depth. Although mylonitic shear zones are only visible to the observer, when uplifted and exposed at the surface, they govern the mechanical behavior of the strongest part of the lithosphere below 10–15 km depth. Mylonitic shear zones dissect plates, thus allowing plate tectonics to develop on the Earth. We review the basic multiscale physics underlying mylonitic, ductile shear zone nucleation, growth and longevity and show that grain size reduction is a symptomatic cause but not necessarily the main reason for localization. We also discuss a framework for analytic and numerical modeling including the effects of thermal–mechanical couplings, thermal-elasticity, the influence of water and void-volatile feedback. The physics of ductile shear zones relies on feedback processes that turn a macroscopically homogeneously deforming body into a heterogeneously slipping solid medium. Positive feedback can amplify strength heterogeneities by cascading through different scales. We define basic, intrinsic length scales of strength heterogeneity such as those associated with plasticity, grain size, fluid-inclusion and thermal diffusion length scale.

For an understanding ductile shear zones we need to consider the energetics of deformation. Shear heating introduces a jerky flow phenomenon potentially accompanied by ductile earthquakes. Additional focusing due to grain size reduction only operates for a narrow parameter range of cooling rates. For the long time scale, deformational energy stored inside the shear zone through plastic dilation or crystallographic- and shape-preferred orientation consumption only a maximum of 10% of energy dissipated in the shear zones but creates structural anisotropy. Shear zones become long-living features with a long-term memory.

A special role is attributed to the presence of water in nominally anhydrous minerals. We show that water directly affects the mechanical equation of state and has the potential to synchronize viscous and plastic flow processes at geological time scale. We have shown that fully coupled finite element calculations, using mechanical data from the laboratory, can reproduce the basic mode of deformation of an entire mylonitic shear zone. The next step of modeling lies in benchmarking basic feedback mechanism in field studies and zooming into the braided network of shear zone structure.

* Corresponding author. Present address: CSIRO Exploration and Mining, 26 Dick Perry Ave., Perth, WA 6151, Australia.
Fax: +61-864368555.
E-mail addresses: klaus.regenauer-lieb@csiro.au (K. Regenauer-Lieb), davey@krissy.msi.umn.edu (D.A. Yuen).
1 Fax: +1-612-6253819.

0012-8252/S - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0012-8252(03)00038-2
without losing the large-scale constraint. Numerical methods capable of fulfilling the goal are emerging. These are adaptive wavelet techniques, and hybrid particle–finite element codes, which can be run over a computational GRID across the net. © 2003 Elsevier Science B.V. All rights reserved.

**Keywords:** plate tectonics; shear localization; mylonitic shear zones; energetics; multiscale modeling

1. **Introduction**

Shear zones in geology occur over many different length scales, from micro (grain size)- to large plate boundary scale (Fig. 1). Plate boundaries define plate tectonics. Shear zones are found on the Earth, Venus, Mars and icy planets such as the Jovian, and Saturnian Moons (Fig. 1). The San Andreas fault zone is an excellent example of a terrestrial large-scale brittle fault zone (Luyendyk and Hornaflus, 1987; Luyendyk et al., 1985; Lyzenga et al., 1986). Shear zone examples with plate scale ductile flow localization are the Alpine Fault in New Zealand (Wellman, 1984), the Kun-Lun and the Altyn-Tagh shear zones in China (Tapponnier and Molnar, 1977). Brittle fault zones can be well traced at depth by narrow seismo-active linea-

---

**Fig. 1.** Shear zones on (a) Venus: Guinevera Planita showing equidistant wrinkles covering large parts of the surface; image scales 40 × 40 km. (b) Europa: ice ridges and grooves forming a criss-cross structure on the Jovian moon (http://www.jpl.nasa.gov/galileo/europa/e4images.html); image scales 1780 × 1780 km. (c) Earth: The San Andreas Fault on the Earth (http://www.scecdc.scec.org/faultmap.html). (d) Microstructural image of a mylonite shear zone: (image scales 4 × 4 cm), http://www.courses.eas.ualberta.ca/eas421/images/photographs/09_9-06qzmylonitegyp.jpg.
ments. Active brittle fault zones can be analysed in terms of their complexity and modes of dynamic interaction by estimates of the fractal dimensions (Matsumoto et al., 1992; Okubo and Aki, 1987), the earthquake statistics (Wyss and Wiemer, 2000) and detailed analyses of the rupture process (Sieh et al., 1993; Zhao and Kanamori, 1993). Ductile fault zones are only accessible when they are exposed to the surface at which point they record a snapshot of the geological history. Indirect observations on active shearing are only available for extreme cases with anomalous heat transfer (Hochstein and Regenauer-Lieb, 1998; Hochstein et al., 1993) and again through secondary observations of the earthquake rupture processes (Wiens and Snider, 2001).

Insight into the physics, dynamics and mechanics of ductile shear zones is, however, imperative for understanding plate tectonics, because plate boundaries are defined by ductile shear zones (Bercovici, 1996, 2002; DeMets et al., 1990). Many approaches have been developed to describe shear zones. The important aspect of time scale has been emphasized in the different fields using various rheologies such as purely viscous on the long time scale to visco-elastic on short time scales. Geodynamic modeling of terrestrial planets uses time scales as defined by cyclic, quasi-periodic behaviour. On the Earth, this is known as the Wilson cycle, on Venus the resurfacing time scale, icy planetary surfaces have a comparatively fast cycle of several hundred years. Earthquake modeling goes down to a shorter time scale with recurrence periods of earthquakes, which are less than 10 ka and the process of the earthquake rupture down to tens of seconds or a few minutes. Most modeling approaches have not considered the coupling of loading rate, the mechanics and energetics of shear zones.

Laboratory analogue models (e.g. Faccenna et al., 1996; Shemenda and Grocholsky, 1994) are obviously limited by availability of materials and laboratory conditions. They fail to reproduce appropriate time scales for the analysis of the delicate influence of coupling thermal diffusion to dynamic gravity loading rate as imposed by thermal expansion and to the shear zone internal temperature sensitive properties. Purely mechanical numerical analyses have been done, e.g. by Buck, Poliakov and Pollitz (visco-elastic but without energy) (Buck and Poliakov, 1998; Poliakov et al., 1994; Pollitz, 2001). We are emphasizing in this article on the nonlinear feedback by considering both thermal–mechanical coupling (the energetics) and using composite rheology, ranging from the simple viscous to complex visco-elasto-plastic rheologies. In doing so, we restrict our description to the depth range deeper than 10 km, because the near surface layers provide additional complexities without contributing much to the strength of the lithosphere. We will highlight the differences between the models and point out where we need to consider dynamical time scales arising in these thermal–mechanical systems. In the sections to come we will describe three kinds of models:

1. Geodynamic modeling: time scale < 500 Ma,
2. Earthquake modeling: time scale < 5 ka,
3. Structural geological modeling no dynamical time scale, only driven by boundary conditions.

Shear zones form as the result of a thermo-mechanical instability that can have many different origins. Prior to the formation of shear zones strain hardening decreases to a critical level which depends on the material, its current state and its $p-T$ condition. The scientific challenges to understanding the dynamics of shear zones are the (sometimes) unobservable dynamics and multiscale physics summarized in Tables 1 and 2. Solutions to the problem of brittle shear zone formation will provide insight into the quasi-periodicity of earthquakes while solutions to the ductile shear zones gives insights into the problem of cyclic-like nature of plate tectonics. This paper focuses on the second problem.

Modern numerical approaches for understanding the dynamics of brittle earthquakes are also discussed. However, we will not go into the details of modelling fault zones in the brittle domain because the multiscale thermal–dynamic material properties are less well constrained than the ductile properties. The brittle strength of the lithosphere is probably overstated. Significant scale dependence of the brittle properties of rocks have been reported in the literature on the brittle field (see e.g. Shimada, 1993). The brittle compressive failure strength of a rock is for instance one order of magnitude smaller at meter scale than at cm scale. Above one meter there appears to be a statistical satisfactory number of planes of weaknesses...
in rocks so that the failure strength does not decrease further. Unfortunately, very huge testing machines are necessary to obtain mechanical data relevant for the larger scale. The necessity for assessing the large-scale has been realized only for the laboratory assessment of friction (Dieterich, 1979a). An equivalent approach is lacking for the compressive failure strength of rocks.

For ductile materials a scale-dependent strength transition is well described between the micro- (microns) and nanoscale (nm). It defines the intrinsic length scale of plasticity (Gao et al., 1999). This scale is more easily accessible to material testing. By analogy to these ductile strain-gradient methods listed in Table 2, we appear to lack in Table 1 a theory that describes the dynamical behavior of brittle material between grain size and meter scale.

Outlining various sections to come, we will in Section 2 summarize the basic equations underlying shear zone formation. In Section 3, we will discuss basic feedback in a one-dimensional shear cell. In

<table>
<thead>
<tr>
<th>Spatial scale</th>
<th>Physics</th>
<th>Input from lower scale</th>
<th>Output to upper scale</th>
<th>Computational methods</th>
<th>Research status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Griffith crack 0.1 – 1 (\mu m)</td>
<td>Atomic bonds</td>
<td>Equilibrium lattice spacing</td>
<td>Effective (damaged) elasticity</td>
<td>Molecular dynamics, finite elements (FE), finite difference (FD)</td>
<td>Understand mechanics and kinetics</td>
</tr>
<tr>
<td>Grain size 1 (\mu m) – 1 cm</td>
<td>Atomic bonds/contact interactions</td>
<td>Cohesive potential across grains</td>
<td>Effective viscosity</td>
<td>Particle dynamics</td>
<td>Link gouge mechanics to fault behaviour</td>
</tr>
<tr>
<td>Fault groups 100 m – 10 km</td>
<td>Coarse grain, planar fault, effective friction</td>
<td>Effective or rate and state variable friction</td>
<td>Effective (\mu), viscosity and elasticity</td>
<td>Finite elements, boundary elements</td>
<td>Understand dynamical modes of faults</td>
</tr>
<tr>
<td>Tectonic plate boundary</td>
<td>Earthquake dynamics</td>
<td>Effective viscoelasticity</td>
<td>No larger scale</td>
<td>Finite elements</td>
<td>Coupling to ductile shear zones</td>
</tr>
</tbody>
</table>

in rocks so that the failure strength does not decrease further. Unfortunately, very huge testing machines are necessary to obtain mechanical data relevant for the larger scale. The necessity for assessing the large-scale has been realized only for the laboratory assessment of friction (Dieterich, 1979a). An equivalent approach is lacking for the compressive failure strength of rocks.

For ductile materials a scale-dependent strength transition is well described between the micro- (microns) and nanoscale (nm). It defines the intrinsic length scale of plasticity (Gao et al., 1999). This scale is more easily accessible to material testing. By analogy to these ductile strain-gradient methods listed in Table 2, we appear to lack in Table 1 a theory that describes the dynamical behavior of brittle material between grain size and meter scale.

Outlining various sections to come, we will in Section 2 summarize the basic equations underlying shear zone formation. In Section 3, we will discuss basic feedback in a one-dimensional shear cell. In
Section 4, we will expand the analysis into a two-dimensional shear zone model and discuss the role of fundamental length scales. These theoretical models are applied in Section 5 to the problem of self-consistent plate tectonics. Section 6 looks into the problem of longevity and memory of shear zones, while Section 7 points out the implications for possible directions in the future of earthquake modeling. We conclude in the summary with a synopsis of length scales obtained from structural field observations and highlight the implications for thermal–mechanical processes inside the shear zone.

2. Mathematical equations

In the following, we will present relations between the stress in a body and its associated cumulative strain (solid mechanics) or strain rate (fluid mechanics). For a comprehensive analysis of shear zones, we cannot restrict ourselves with a one-dimensional analysis as shown in Fig. 2, but we need to go to a full three-dimensional formulation. We first define the basic quantities. The stress matrix describes the traction being carried per unit area by any internal surface in the body under consideration. This is the “Cauchy stress”, which is given by

\[
\begin{bmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} \\
\sigma_{31} & \sigma_{32} & \sigma_{33}
\end{bmatrix}
\]

where \(\sigma_{ij}\) is the force per area acting on surfaces facing in the \(i\)-direction and pulling/pushing it in the \(j\)-direction. We will always imply the so-called “Einstein convention” (Hill, 1950), i.e. a summation over repeated indices. The Cauchy stress gives the “true” stress for any particular choice of orientation of the coordinate system. It is useful to derive independent invariants from the Cauchy stress:

\[
p = \frac{1}{3} \left( \sigma_{ik} \delta_{ij} \right) = \frac{1}{3} (\sigma_1 + \sigma_2 + \sigma_3)
\]

where \(p\) is the force per area acting on surfaces facing in the \(i\)-direction and pulling/pushing it in the \(j\)-direction. We will always imply the so-called “Einstein convention” (Hill, 1950), i.e. a summation over repeated indices. The Cauchy stress gives the “true” stress for any particular choice of orientation of the coordinate system. It is useful to derive independent invariants from the Cauchy stress:

\[
p = \frac{1}{3} \left( \sigma_{ik} \delta_{ij} \right) = \frac{1}{3} (\sigma_1 + \sigma_2 + \sigma_3)
\]

This is also known as the trace, first invariant or isotropic part of the stress tensor. \(\delta_{ij}\) is the Kronecker

![Fig. 2. Average macroscopic shear stress–strain diagram as recorded by the gauges of a laboratory experiment showing typical conditions leading to the development of a shear zone after critical strain hardening for different strain rates. Strain hardening must not at all cost be negative for the onset of a shear localization but can be positive for a variety of materials, if the confining pressure is low. At steady state, the shear zone has reached its maximum width. The 1D shear zone shows the mathematical idealization with viscous, visco-elastic, elasto-plastic or elasto-visco-plastic rheology.](image)
delta, which is one when the indices are equal and zero for unequal indices. We will use the term “pressure” \( p \). Most flow laws are independent of pressure so that it is convenient to define the flow law on the basis of the square root of the second invariant of the deviatoric stress tensor which itself is defined by subtracting the pressure from the Cauchy stress \( \sigma'_{ij} = \sigma_{ij} - p \).

Effective Stress \( \sigma' = \sqrt{\frac{1}{2} \sigma'_{ij} \sigma'_{ij}} \)  

This invariant of the deviatoric stress tensor is also known as the “root mean square”, or the “effective shear” stress, which is a scalar. Norms differing by a factor can be found in the literature (Chakrabarty, 2000; Ranalli, 1995). In Appendix A, we discuss and explain the alternative definition of the effective stress, which is more convenient for implementing experimental flow laws into numerical models. In Appendix B, we show how to turn the scalar effective stress–strain rate relation into tensorial flow laws.

For the choice of definition of strain and strain rate, it is necessary to consider energy dissipated by the deformation (Fig. 3), i.e. we want to define a strain that when multiplying its increment \( d\varepsilon_{ij} \) with the Cauchy stress gives the work done in the unit body. This collapses to a one-dimensional case to the familiar definition of work by force times displacement. In Fig. 3, we show the case of “associated plasticity”, i.e. co-axial stress and strain increment tensors (Appendix B). This is a necessary condition when plastic deformation does not imply volume (surface energy) or

![Fig. 3. The second invariant of the deviatoric stress tensor traces a cylinder around the first invariant \( p \) when visualized in the principal stress space (principal stresses are the axes of a 3D Cartesian space). In the ideal rigid-plastic case, the cylinder defines the von Mises Yield Envelope (Chakrabarty, 2000). The inside of the cylinder gives an elastic stress state (in this section \( E = \infty \), i.e. rigid, we will relax this assumption later), while any plastic strain is possible when the cylinder is reached. In the general plastic case, the cylinder is allowed to swell (strain harden, see Eq. (12)) or contract (strain soften to be discussed later) as a function of the plastic strain. In the absence of dilatancy, stress and strain tensors must be conjugate as indicated by the vectors in the deviatoric stress space (white circle going through the origin also known as the \( \pi \)-plane), i.e. the strain increment vector must always be normal to the yield surface (plastic normality rule). For the case of a viscous flow rule, the initial cylinder is lacking (except for the Bingham solid) but conjugacy between stress and strain rate still holds. The second superposed coordinate system, the incremental principal strain, must be replaced in the viscous case by the principal strain rates.](image-url)
other changes of energy. The infinitesimal strain increment tensor for time $d\tau$ is

\[
\text{Strain increment and strain rate}
\]

\[
d\varepsilon_{ij} = \frac{1}{2} \left( \frac{\partial dX_i}{\partial x_j} + \frac{\partial dX_j}{\partial x_i} \right), \quad \dot{\varepsilon}_{ij} = \frac{d\varepsilon_{ij}}{d\tau}
\]  

(4)

where $dX_i$ is the infinitesimal displacement of a particle in time $d\tau$ with a current position vector $x_i$. Note that, in linear elasticity, it is customary to omit the increment and use the above definition as a small strain measure $\varepsilon_{ij}$. In plasticity, an appropriate integration is mandatory unless proportional straining is assumed. Analogous to the definition of invariants for stress we define the roots of the second strain increment and strain rate invariants as “effective” strain increment or strain rate:

\[
\text{Effective strain and strain rate}
\]

\[
d\varepsilon = \sqrt{\frac{1}{2} d\varepsilon_{ij} d\varepsilon_{ij}}, \quad \dot{\varepsilon} = \sqrt{\frac{1}{2} \dot{\varepsilon}_{ij} \dot{\varepsilon}_{ij}}.
\]  

(5)

2.1. No elasticity and compressibility

Consider slow motion. We neglect inertial forces and describe a balance of all forces in a unit volume by

\[
\text{Momentum conservation}
\]

\[
d\sigma_{ij} + B_i = 0
\]  

(6)

where $B_i$ is the sum of the body forces. We introduce the plastic flow rule of a solid and the viscous flow rule of a fluid and show their relation. In plasticity or fluid dynamics, the strain increment or strain rate, respectively, are given by a function of the effective shear stress

\[
\text{Plastic and Viscous Flow Rule}
\]

\[
d\varepsilon = f(\sigma'), \quad \dot{\varepsilon} = f(\sigma')
\]  

(7)

However, plasticity deals with the stress-strain relation, while the strength of fluids is described by the strain rate. We point out here that the plastic flow rule is dimensionally consistent, i.e. time does not enter in the equations but it certainly appears in the viscous flow rule (Hill, 1950). Associated flow implies that the strain increment or the strain rate is everywhere normal to the flow potential (Fig. 3) whether there exists a finite yield surface (plasticity) or a continuous potential (viscous flow).

In the following nomenclature, we will separate plasticity-based formulations as defined by the first part of Eq. (7), usually describing a yield phenomenon attributed to the dislocation controlled flow, from fluid dynamic approaches, mostly characterized by diffusion without a yield phenomenon, by the second part of Eq. (7). Plasticity often involves the consideration of elasticity while in fluid dynamics elasticity is frequently neglected. We will, however, also discuss hybrid rheologies that comprise elasticity and viscosity, or elasticity and plasticity and present results for the complete rheological elasto-visco-plastic approach.

Examples for visco-plastic flow rules are:

**Bingham’s Viscous Flow Rule**

\[
\dot{\varepsilon} = f(\sigma') = \begin{cases} 
\dot{\varepsilon} = 0, & \sigma' < \sigma_y \\
\dot{\varepsilon} = \frac{1}{2\eta_c}(\sigma' - \sigma_y), & \sigma' \geq \sigma_y
\end{cases}
\]  

(8)

where $\eta_c$ is the effective viscosity. The Bingham solid incorporates plasticity into the standard Newtonian viscous flow rule. Newtonian viscous flow without plasticity is recovered when the yield stress $\sigma_y=0$ (Fig. 4).

**Power law Viscous Flow Rule**

\[
\dot{\varepsilon} = f(\sigma') = a^{-n}(\sigma')^n
\]  

(9)

For modeling the lithosphere, this power-law is often used with various values of exponent $n$ (Chester, 1995; Christensen, 1992; Lenardic et al., 1995). It varies between $n=1$ for standard Newtonian viscous flow ($a=2\eta_c$), over $n=3–4.5$ from laboratory data,
and to a very high \( n \) \((n = 35)\) to reflect a pseudoplastic flow law (Fig. 5). For \( n \to \infty \), we recover the ideal rigid-plastic flow law with a constant yield stress \( \sigma' = \sigma_y = a \) for any strain rate (Nye, 1953). A third way of incorporating plasticity into viscous flow is the exponential flow law.

Ludwik’s dynamic plasticity law

\[
\dot{\varepsilon} = \dot{\varepsilon}_0 \exp \left( \frac{\sigma' - \sigma_0}{b} \right) 
\]

solved for stress \( \sigma' = \sigma_0 + b \ln \left( \frac{\dot{\varepsilon}}{\dot{\varepsilon}_0} \right) \), \( \dot{\varepsilon} \geq \dot{\varepsilon}_0 \) \((11)\)

where \( \sigma_0, b \) and \( \dot{\varepsilon}_0 \) are material constants. The “over-stress” formulation of this flow law is obvious, when the flow law is solved for stress, giving the logarithmic Eq. \((11)\). Here, \( \sigma_0 \) is the initial yield stress associated with a characteristic strain rate \( \dot{\varepsilon}_0 \) for the onset of creep. For higher strain rates, the stress increases by an increment that is controlled by the logarithmic term.

Up to now, we have discussed extensions of viscous flow theory, thus allowing incorporation of ideal plasticity into the viscous flow. In the classical Newtonian viscous case viscosity is the only material parameter introducing time dependence in the flow rule. For incorporating plasticity, at least one additional material parameter is necessary for scaling accurately the yield stress.

Classical plasticity uses the same Bingham style flow rule, but is independent of time. Assuming proportional straining, the integrated strains replace the incremental strain and we can come up with an analogous equation.

Rigid Plastic Flow Rule

\[
\varepsilon = f(\sigma') = \begin{cases} 
0, & \sigma' < \sigma_y \\
\dot{\varepsilon} = c^{-n}(\sigma')^{-n}, & \sigma' \geq \sigma_y 
\end{cases} 
\]

This is the popular power-law hardening law (Ludwik, 1909) that describes strain (work) hardening by two material parameters the constant stress \( c \) and dimensionless \( n \). This particular form of strain hardening can be derived from the theory of defects (Hirsch, 1975). It describes the increasing strength of a crystalline solid owing to an increase in dislocation density as the strain increases. Again, if \( n \to \infty \), we recover the ideal rigid-plastic body where the plastic stress does not increase with strain.

In sum, we can now come up with an extension to Ludwik’s visco-plastic formulation with strain hardening/softening.

Generalized dynamic plasticity law

\[
\sigma' = f(\varepsilon) + b(\varepsilon)f(\dot{\varepsilon}) 
\]

Now, the yield stress \( f(\varepsilon) \) and the stress scale factor \( b(\varepsilon) \) of the flow term are also a function of plastic strain. Eqs. \((1)–(13)\) provide a complete set for solving the momentum-rheology equations.

2.2. Add temperature and pressure but without water

When considering temperature in addition, we have to solve for the energy equation. Using a Lagrangian framework, i.e. the equation is solved with reference to a moving particle indicated here by the substantial derivative \( D T/Dt \) in which case the advection of heat drops out of the Eulerian equation. In considering temperature as the thermodynamic variable (Appendix C), we obtain

Temperature Equation

\[
\rho C_p \frac{DT}{Dt} = \chi \sigma' \dot{e}_{ij} - \kappa \rho C_p \nabla^2 T + \sum \Omega_i 
\]

where \( \kappa \) is the diffusivity, \( \rho \) the density, \( C_p \) the specific heat, \( \chi \) the mechanical heat conversion efficiency and \( \Omega_i \) the sum of additional heat sources such
as radiogenic heating, heat of solution, phase transition, chemical reaction and other terms discussed in the chapter on volatiles and equation of state.

In the theory of thermally activated creep, temperature and pressure enter through an exponential thermodynamic term with Arrhenius dependence

\[
\dot{\varepsilon} = f'(\sigma')\exp\left(-\frac{Q + pV}{RT}\right)
\]  

(15)

where \(Q\) is the activation energy of the particular flow creep mechanism, \(V\) the corresponding activation volume and \(R\) the universal gas constant. The Arrhenius term in the exponential is additionally controlled by the presence of water.

2.3. Add grain size

Up to now, we have given equations for creep based on line defects, i.e. dislocation glide (the Peierls stress mechanism, a Ludwik’s law style flow law) and dislocation climb (power-law) processes. Plasticity by line defects gives mixed plastic and viscous constitutio

\[
\dot{\varepsilon}^D = \dot{\varepsilon}_0 \left(\frac{g_0}{g}\right)^m \sigma^s \exp\left(-\frac{Q^D + pV^D}{RT}\right)
\]  

(16)

where \(g_0\) is the initial grain size, \(g\) the current grain size and \(m\) the grain size exponent (between 0.3 and 0.5) (Van Swygenhoven, 2002) and \(s\) the stress exponent which is close to unity (Mei and Kohlstedt, 2000). Positive feedback comes in during the deformation through strain-dependent grain size reduction given by

\[
\frac{dg}{dt} = \lambda \dot{\varepsilon}^D (g_t - g) + K_0 \exp\left(-\frac{H}{RT}\right)
\]  

(17)

where \(\lambda\), \(K_0\) and \(H\) are material constants. The first term on the right-hand side describes the effect of grain size reduction and the second term describes the effect of grain growth (Karato, 1989).

2.4. Add water

Water changes the rheology in several ways. We first discuss the effect when only minor quantities of water are added to nominally anhydrous minerals, i.e. the rock incorporates water into the solid without microstructural modification. In this case water has two major effects.

Water weakening

\[
\dot{\varepsilon} = \alpha f'(\sigma')\exp\left(-\frac{Q^* + pV^*}{RT}\right)
\]  

(18)

Water changes the activation volume and the activation energy in the Arrhenius term because of the formation of new hydroxyl bonds. It accelerates creep rates by a scalar factor \(\alpha\), which ranges between 0.1 and unity depending on water content (Jung and Karato, 2001). The activation energy \(Q^*\) is lower than the dry value but only by about 10%. The activation volume term \(V^*\) that is attributed to water changes its pressure sensitivity. Large variations up to 50% are reported in the literature. Its value is difficult to determine with precision, but the general magnitude will give a sense of the physics.

Since water has the same principal effect on all creep mechanisms, it is most prominent in highly non-linear flow laws, especially where it has a rather strong influence on degree of non-linearity. The power-law has already been introduced as a non-linear flow law. The pre-exponential weakening factor \(\alpha\) linearly scales the magnitude of the yield like transition from high viscosity at low strain rates to low viscosity at high strain rates. However, the sharpness of the transition is not affected since it is only controlled by the exponent \(n\). For high water content, the overall weakening through the addition of water can reach an order of magnitude. For rocks the power-law, exponent never exceeds \(n = 5\), so we cannot call this a true yield phenomenon as in the pseudo-plastic case. Flow laws, where water has a strong effect on the yield stress, should indeed play a prominent role in the nucleation of shear zones. Water indeed has a fundamental influence on the exponential flow law that has been applied to indentation hardness experi-
ments of quartz and olivine (Evans, 1984; Evans and Goetze, 1979; Goetze and Evans, 1979). It is also known as “low temperature plasticity”, “Peierls stress” or “Dorn-Harper” creep law.

Peierls strain–stress law

\[
\dot{\varepsilon}_L = \dot{\varepsilon}_0 \exp \left( - \frac{Q_L^* + pV_L^*}{RT} \left( 1 - \frac{\sigma'}{\sigma_0} \right)^2 \right) \quad (19)
\]

This law is more complex than Ludwik’s dynamic plasticity law (Eq. (11)) in that it has also an additional sharp transition at high stress. We show here that water has an influence on its first embedded yield criterion. Inverting Eq. (19), we obtain

Peierls stress–strain law

\[
\sigma' = \sigma_0 - \sigma_0 \sqrt{\left( - \frac{RT}{Q_L^* + pV_L^*} \right) \ln \left( \frac{\dot{\varepsilon}_L}{\dot{\varepsilon}_0} \right)}, \quad \dot{\varepsilon}_L > \dot{\varepsilon}_0 \\
\]

(20)

This flow law recovers the ideal plastic case for a hypothetical \( T = 0 \) K when the second term vanishes and \( \sigma' = \sigma_0 = \sigma_y \) is the ideal yield stress, the so-called “Peierls stress”, for any strain rate. The reference strain rate \( \dot{\varepsilon}_0 \) is a material constant. The thermally activated second term influences the yield phenomenon and yielding occurs with a characteristic strain rate \( \dot{\varepsilon}_0^* \) (Branlund et al., 2001; Regenauer-Lieb et al., 2001).

Characteristic strain rate after yield

\[
\dot{\varepsilon}_0^* = n \dot{\varepsilon}_0 \exp \left( - \frac{Q_L^* + pV_L^*}{RT} \right) \\
\]

(21)

In analogy to the discussion of Ludwik’s style dynamic plasticity law (Eq. (11)), we now obtain a characteristic strain rate (Fig. 4) for the onset of creep that is no longer a material constant but depends on pressure, temperature and water content. Consequently, the initial yield stress also depends on these thermodynamic quantities and the water content.

We have thus far given given constitutive equations for dynamic visco-plasticity for the case of low water content. If volatiles in excess of their solubility are present in the solid, volatiles will act as a damaging agent, i.e. they will create voids (fluid inclusions), which weaken the rock matrix as considered by the strain-dependent parameter in the generalized dynamic plasticity formulation. In Section 2.5, we will describe a self-consistent strain-weakening, theory arising from influx of volatiles.

2.5. Add volatiles

The von Mises yield envelope shown in Fig. 3 is insensitive to the pressure. The lithostatic pressure ensures that the mechanically strong part of the lithosphere is in an overall compressive regime. Therefore, the von Mises envelope is a safe approach for large overburden pressure and for the case of absence of volatiles. However, when volatiles are present, the fluid pressure compensates the overburden around the inclusion. For simplicity, we will assume that the fluid pressure is lithostatic, if there is no other load than a pure lithostatic load. This implies that tensile stress states are possible locally around the fluid under the addition of a tectonic load (Petrini and Podladchikov, 2000).

When dealing with this problem mathematically we have to bear in mind that the volatile content of a deep seated rock is not more than 0.5 wt.% equivalent to 3% of volume of the intact rock. This is the largest content of fluid inclusions reported in the literature (Roeder, 1965). We therefore cannot imply a macroscopic brittle failure criterion on the basis of a global effective stress principle. Instead, we have to define the local void volume as a fracture controlling parameter. If we define the relative density of the rock by \( r \), which is the ratio of solid over the total volume, then we obtain the void volume fraction as

\[
\text{Void Volume Fraction } f = 1 - r \\
\]

(22)

Because of the small void volume fraction, we can only assume at a depth, say greater than 10 km, that the bulk of the rock matrix is still deforming by crystalline plasticity. The yield criterion therefore is still based on the von Mises Criterion with an extension for pressure sensitivity and void volume fraction. A hyperbolic cosine surface has been suggested (Tvergaard, 1987) that truncates the von Mises style
envelope in the tensile domain as a function of the void volume fraction.

\[
\frac{(\sigma')^2}{(\sigma_y)^2} + 2q_1 f \cosh \left( q_2 \sqrt{3} \frac{P}{\sigma_y} \right) - \left( 1 + q_3 f^2 \right) = 0
\]

(23)

where the first term is the classical von Mises yield envelope, the second term is important for tensile stress states and the third term tracks the damage created by the voids, \(q_{1,2,3}\) are material parameters. The yield envelope is shown in Fig. 6. The dynamic evolution of volatiles is given by the sum of the nucleation rate of new voids plus the growth rate of existing voids denoted by subscripts

Dynamic void evolution \( \dot{f} = \dot{f}_{\text{nuc}} + \dot{f}_{\text{gr}} \)

(24)

where the growth rate is controlled by the mass conservation

Void growth rate \( \dot{f}_{\text{gr}} = \dot{\varepsilon}(1 - f) \)

(25)

and the nucleation is assumed to be either strain rate controlled

Ductile void nucleation rate \( \dot{f}_{\text{nuc}} = A \dot{\varepsilon}^h \)

(26)

or stress rate controlled

Brittle void nucleation rate \( \dot{f}_{\text{nuc}} = B(\dot{\sigma}' + \dot{\rho}) \)

(27)

where \(A\) and \(B\) are normal distributions around the nucleation stress or strain for plastic strain (ductile) or plastic stress (brittle) controlled nucleation (Needleman, 1994) and \(\dot{\varepsilon}^h\) describes the corresponding hardening strain rates. Both Eqs. (26) and (27) describe material specific energy density rates for nucleation of voids. In the ductile case, voids are nucleated due to dislocation climb and glide while in the brittle theory they are nucleated by classical cleavage cracks. Expressions for accelerated growth after reaching a critical void volume fraction can be found in the cited literature (Dodd and Baiy, 1987).

We note here that the last three equations also feed back into the energy Eq. (14) through the shear heating
efficiency $\chi$. The generation of new surface energy stores some fraction of the deformational work to be converted into heat later, thus lowering $\chi$. However, the low void volume ratio considered here, implies that other structural defects such as energy absorbed into increasing the dislocation density are dominant. Therefore, in solid mechanics, it is common practice to obtain $\chi$ from experiments such as thermographic imaging techniques (Chrysochoos and Belmahjoub, 1992) and consider the stored energy fraction $1 - \chi$ by the pre-factor $\chi$ (see also Appendix C). For large strain, the shear heating efficiency $\chi$ of most materials is found to lie between 85% and 95%. This means that almost all of the deformational work is converted into heat and very little is stored in microstructural processes.

A very similar fluid mechanical approach in two-phase systems has been developed recently (Bercovici and Ricard, 2003; Bercovici et al., 2001a,b; Ricard et al., 2001), which has the advantage of considering fully the thermodynamic work including explicit terms for void volume fraction (Eqs. (65) and (66), Bercovici et al., 2001a). Being a viscous approach, it can only describe long timescale processes. This approach ignores, however, the duality in the physics of void creation as it is expressed in Eqs. (26) and (27). At present, the theory of Bercovici and Ricard is tuned to the brittle top 10 km where damage is accommodated by brittle micro-cracking.

2.6. Add equation of state and compressibility

The constitutional laws laid out above have been formulated for a rigid-plastic, incompressible viscous medium only defining its deviatoric strength. Although the equation of state has already been used implicitly in the flow laws, an important element of physics is missing. Incompressible media are mathematical idealizations, not to be found in nature. We start here with the ideal gas, which has no infinite compressible strength, its density being described by the equation of state. Avogadro showed that under the same $p$–$T$ conditions the number of molecules in a given volume is constant.

Ideal Gas Equation of State

$$ p = \frac{n}{V} RT = \rho_{mol} RT $$

(28)

where $n$ is the number of mol in the volume $V$ (we recall 1 mol is defined on the basis of the quantity of carbon isotopes contained in 12 g of $^{12}$C which is $6.022136 \times 10^{23}$ mol$^{-1}$). The ideal gas constant $R$ scales the proportionality between thermal pressure and internal energy. Defining the molar density $\rho_{mol}$ as the inverse of the molar volume, the equation of state of water is formulated (Pitzer and Sterner, 1994). Water has no deviatoric strength and the non-ideal Helmholtz free energy (see Appendix C) with the addition of the ideal gas term of water is of following type

Water equation of state

$$ p = RT \left( \rho_{mol} + C_i(T) \rho_{mol}^2 + \sum c_{i(T)} \Theta_i \right) $$

(29)

where $\Theta_i$ contains higher orders of $\rho_{mol}$ and temperature sensitive coefficients $c_{i(T)}$ as well as two exponential terms necessary near critical point of water. For a general solid with less extreme property changes, similar expressions can be found (Dorogokupets, 2001; Poirier and Tarantola, 1998).

We are now in a position to reconsider the energetics of compressibility. In the energy Eq. (14), we have only looked at work done under deviatoric stress. Compressibility introduces volume changes which are recoverable upon unloading. An additional recoverable term related to the elastic volume change appears in the energy equation (see Appendix C).

$$ \frac{\rho C_p}{\rho} \frac{D T}{D t} = \chi \frac{\dot{\epsilon}_{ij}}{\dot{\epsilon}} \dot{\epsilon}_{ij} + \lambda_{th} T_{eqv} \frac{D p}{D t} - \rho C_p \nabla^2 T + \sum \Omega_i $$

(30)

where $\lambda_{th}$ is the thermal expansivity which multiplied by the adiabatic temperature change $T_{eqv}$ describes the recoverable elastic volume change. The energy equation and the equation of state are coupled equations for pressure and internal energy (Appendix C). Simultaneously solving for the equation of state, the rheology and the energy equations is a fundamental prerequisite to understanding the complete thermo-mechanical structure of shear zones.

2.7. Additive strain rate approximation

We have separated out the energetics of deformation into a conservative and dissipative component. Elastic work is stored reversibly as strain energy while
plastic/viscous work is either dissipated immediately as heat through viscous processes, or stored as structural defects. In addition, there are important differences in the intrinsic length and time scales of elastic and plastic and viscous deformation. Visco-plastic deformation is transmitted by the motion of line defects, so there is a rate limited process controlled by atomic relaxation times. Elastic strain relies on electromagnetic waves, so it is determined by electronic relaxation times. The length scale of plastic deformation relies on the size of line defects and magnitude of lattice vibrations. The length scale of viscous deformation relies on thermal and chemical diffusion through lattice and crystal sizes, while elastic deformation only relies on electronic (ionic or covalent) potential in a perturbative sense. Since these processes are fundamentally different and operate separately, a single elasto-visco-plastic constitutive equation does not exist. Instead, we use the additive strain rate decomposition.

\[
\dot{\varepsilon}_{ij} = \dot{\varepsilon}_{ij}^{E} + \dot{\varepsilon}_{ij}^{T} + \dot{\varepsilon}_{ij}^{P} + \dot{\varepsilon}_{ij}^{D} \quad (31)
\]

where the total strain rate is a composite of elastic, thermal (appearing through the recoverable thermal expansion), Peierls, power-law and diffusion creep strain rates are identified by their superscripts. Also inherent in this assumption is that the deviatoric strength, not defined in the equation of state, can be added on the basis of the deviatoric properties of creep, plasticity and elasticity laws. This composite rheological law is the current continuum mechanics approach to thermodynamics in the deviatoric stress space. Recent non-equilibrium molecular dynamics calculations have lent strong support (Holian and Lomdahl, 1998) to these macroscopic simplifications.

2.8. Localization mechanisms

2.8.1. Constitutive theory

In geodynamics, it is common practice to consider temperature and pressure only through their immediate effect on the strength parameters, without considering a fully coupled feedback (see Yuen et al., 2000). We will refer to these approaches as mechanical models. More elaborate approaches consider heat conduction (second term in Eq. (14)) and radioactive heat generation. Then, it suffices to calculate a thermal solution say every several million years because of the slow pace of conduction and radioactive heat generation. Such approaches are sometimes called thermo-mechanical (Beaumont et al., 1996b). For the topic discussed here, this is a misnomer. A staggered momentum and thermal solution does not have the potential for thermal feedback. Hence, staggered solutions must nucleate shear zones the same way the mechanical models do. For the purpose of nucleation shear zones, they also belong to the mechanical group implying that flow localization is entirely characterized by temperature sensitive-constitutive laws.

The mechanical way of solving the problem is basically thermally decoupled and it is necessary to nucleate shear zones through the feedback between momentum and rheology only (Poliakov and Herrmann, 1994; Tommasi et al., 1995). It is amenable to full theoretical assessment if the rate effects are also neglected (Rice, 1977). In this approach shear zone, formation is understood as an instability that can be predicted from the pre-localization constitutive equations.

Conditions are sought at which some small perturbation is allowed accelerated growth so that initial uniform smoothly varying deformation turns into highly localized deformation, a flow bifurcation occurs. Shear zones form on potential shear planes if the strain hardening on those potential planes is lower than a critical strain hardening $h_{crit}$ depending on rheology (Fig. 2). While the actual hardening is a function of the deformational history, it is possible to predict the tendency to spontaneous localization by the magnitude of the hardening parameter for specific rheologies, which turns out to be a material parameter.

In earlier papers on localization (see Needleman and Tvergaard, 1992; Rice, 1977 for reviews), much work has been devoted to theoretical ends for understanding constitutive instability (Appendix B). The Earth’s brittle crust localizes readily in the form of brittle shear zones, being entirely described by the constitutive theory. However, according to the same theory, the ductile part of the lithosphere cannot localize. In mechanical models, the ductile part of
the lithosphere can only be deformed by way of homogenous shear. Mechanical models therefore have to overemphasize the role of the brittle part of the lithosphere or fail to model discrete plate boundaries accommodated by ductile, mylonitic shear zones.

Additional weakening in the ductile level can be obtained by treating also the energy equation, which is neglected in the constitutive theory. We will therefore refer to this relatively new theory as the “energy theory” of localization. In a first step, following the arguments of Hobbs and others (Backofen, 1972; Hobbs et al., 1986), the theory for constitutive instability can, however, be recast to explore basic conditions for energetic instability without solving the energy equation. We will use this extended constitutive theory as a lead in for reviewing the energy theory of localization.

For the ductile flow, laws defined above potential planes of localization are defined by the second invariant of the deviatoric stress tensor and we can write a suitable criterion for localization based on effective scalar quantities (Appendix A)

\[ \frac{d\sigma'}{d\epsilon} \leq h_{\text{crit}} \]  \( (32) \)

A very similar formulation can be cast for the nucleation of compaction or dilatation bands, when replacing the effective shear stress by the pressure and the effective shear strain increment by the volumetric strain. For the high-pressure conditions, in geodynamic problems, we do not need to consider such pressure-dependent localization phenomena. However, it turns out that \( h_{\text{crit}} \) is very sensitive to the deviations from the smooth co-axial von Mises style yield surface assumed so far. Geological materials in the top 10 km can be described by a yield surface that has corners (Coulomb envelope) and has non-coaxial flow (strain-increment and stress vectors are not collinear as in Fig. 3, see also Appendix B). Both factors promote shear zones but especially the latter ensures that the material localizes readily (Poliakov et al., 1994). For the incompressible, non-coaxial case the constitutive theory (Rudnicki and Rice, 1975) predicts that bifurcations can occur for any amount of strain hardening. At greater depth in the lithosphere, crystalline plasticity applies and flow becomes co-axial. In the absence of yield vertices and non-coaxial flow (Appendix B), the criterion for onset of instability is negative strain hardening \( h_{\text{crit}} < 0 \) in the time increment \( dt \), i.e. the rock must become weaker with deformation for shear zones to nucleate spontaneously.

We have already discussed the void-volatile feedback mechanism as a potential mechanism for strain weakening in the co-axial domain. It has been argued that void coalescence can also lead to departures from co-axiality, hence enhancing the tendency for localization (Rice, 1977) according to the constitutive theory (Appendix B). Propagating void sheets may also significantly change the energetics of deformation and localize as discussed in the energy theory. This leads to the appearance of additional, destabilizing (negative) terms in the hardening law. A preliminary assessment of these additional terms is that propagating void sheets act destabilizing so that a conservative assumption is to equate the additional terms to zero (Hobbs et al., 1986).

Eq. (32) can be extended for the rate-sensitive material as a suitable criterion for shear zone formation.

\[ \frac{d\sigma'}{d\epsilon} = \frac{\partial \sigma'}{\partial \epsilon} + \frac{\partial \sigma'}{\partial \dot\epsilon} \frac{d\dot\epsilon}{d\epsilon} \leq h_{\text{crit}} \]  \( (33) \)

Investigating the fluid dynamic visco-plastic formulations laid out above, we find that they have zero strain hardening (first partial derivate drops out) so they should be good candidates for spontaneous shear zone nucleation. However, all of these flow laws have a positive strain rate hardening or zero strain hardening in steady state. In a time-dependent circumstance, the viscosity drops when creep is accelerated but the stress still increases. By analogy to the rate-independent solid, we expect that shear zones do not form spontaneously in a smoothly varying deformation field. Shear zones can still form for strongly heterogeneous boundary conditions or for negative strain rate hardening through spontaneous transitions from one flow law to another weaker form.

An example for such a transition is shear zone formation owing to grain size-sensitive creep from
mechanical standpoint (Braun et al., 1999). Dynamic recrystallization relies on subgrain formation during the movement of line defects in the power-law regime. When the grain size is reduced sufficiently, a switch in flow law can occur to a weaker grain size diffusion dominated Newtonian viscous flow. At this point, strain rate hardening can become negative and shear zones can nucleate spontaneously. This transition is, however, also thermally activated (Kameyama et al., 1997) and a full thermal–mechanical energy assessment is therefore necessary.

Leaving now the mechanical approaches and turning over to the thermo-mechanical shear zone models, we have now to consider the energetics in a more self-consistent manner. The shear heating term in the energy equation and the thermal expansivity both feed back positively into the momentum-rheology equations. It is noteworthy that for increasing strain or strain rate hardening shear heating also increases, i.e. the vigor of thermal feedback increases. Therefore, for most materials, the contribution of temperature in the hardening law is negative (Backofen, 1972) thus promoting the spontaneous nucleation of shear zones.

Thermal – rheological weakening
\[
\frac{d\sigma'}{d\varepsilon} = \frac{\partial \sigma'}{\partial \varepsilon} + \frac{\partial \sigma'}{\partial \dot{\varepsilon}} \frac{d\dot{\varepsilon}}{d\varepsilon} + \frac{\partial \sigma'}{\partial T} \frac{dT}{d\varepsilon} \leq h_{\text{crit}}
\] (34)

Thermal–mechanical feedback is dependent on the thermal scaling length introduced in Table 2. If the deformation is fast (high strain rates), we obtain a short scaling length on the order of the shear zone thickness and we can discuss this in terms of near-adiabatic conditions. For slow processes, heat conduction plays a role. Heat conduction is a negative feedback process and it will consequently retard the onset of thermo-mechanical flow bifurcation, or it will inhibit this kind of localization phenomena altogether. Therefore, a fully coupled...
solution including momentum, rheology and energy equations is necessary to comprehensively assess the potential of generating shear zones at greater depth.

2.8.2. Energy theory

The physics underlying ductile faulting can be understood on the basis of a feedback diagram in Fig. 7. In the constitutive theory, we only consider the simpler feedback between the rheological and momentum equations alone, with pressure $p$ being the feedback variable. This is sufficient for modeling faults in the brittle field. The energy theory explains faulting in the ductile level and considers at least one additional feedback variable, which is either temperature $T$ or new surface energy by ductile cracking.

The energy theory of shear zone formation and equilibration is the core subject of this review. It is the key to understand ductile, mylonitic shear zones. Its geo-scientific implementation is opposed to the classical constitutive, purely mechanical theory in the subsequent chapters, where the weaknesses of the latter are resolved in the subheadings on energy theory. Energy-based approaches to localization have been developed independently in the engineering and the geodynamics community and have been rapidly evolving over the recent years. The energy hypothesis for shear zone formation has been put forward as a rigorous theory (Appendix C) in the solid mechanical engineering literature around 10 years ago (Cherukuri and Shawki, 1995a,b; Sherif and Shawki, 1992; Zbib, 1992), while it has emerged in approaches to localization phenomena in fluid flow much earlier (Gruntfest, 1963). The theory has originally been restricted to describe thermal feedback only, but recent extensions also include the effect of surface energy (Bercovici and Ricard, 2003; Lyakhovsky, 1997; Regenauer-Lieb, 1999).

Unfortunately, in putting forward a new energy theory, which is still far from being complete, there has been little exchange between engineering and geophysical approaches. This review focuses on the recent advances in geosciences and therefore gives a somewhat biased view. It is beyond the scope of this review to unify the advancements, however, whereever possible proper credit to the engineering community is given (such as in Appendices B and C).

3. 1D shear zones

One-dimensional analyses have the advantages of extremely high resolution, like hundreds to millions of grid points, and provide an estimate of time scale. A problem is of course the choice of appropriate boundary conditions. Assuming to a zeroth order a shear cell as shown in the inset of Fig. 2. We will discuss the fluid dynamic approaches first, then proceed to the visco-elastic and finally the elasto-plastic case studies.

3.1. Viscous thermal–mechanical feedback

A number of approaches have been formulated in the 1980s that solve for fluid dynamic shear zones by thermal–mechanical feedback in Newtonian and power-law fluids chosen as a proxy for lithosphere–asthenosphere conditions. Two end member boundary conditions for the one-dimensional shear zone model have been assumed: one in which the shear zone is driven by constant velocity, the other in which the shear zone is driven by a constant shear force (Fleitout and Froidevaux, 1980; Locket and Kusznir, 1982; Schubert et al., 1976; Schubert and Yuen, 1977, 1978; Yuen et al., 1978; Yuen and Schubert, 1977). The equation of momentum conservation (Eq. (6)), the power-law (Eq. (8)) and the energy Eq. (14) are the only ingredients of the analysis. A small thermal, compositional or velocity perturbation is assumed to analyze the stability of basic shear flow. Since no strain hardening was assumed it was found that shear zones form readily under both boundary conditions. Constant stress boundary conditions have been repeatedly found to lead to self-feeding thermal runaway instabilities if the shear stress is high enough (Melosh, 1976; Spohn, 1980). Obviously, constant stress boundary conditions require special settings in an Earth-like scenario. We prefer to discuss runaway instabilities later using a choice of more realistic lithosphere rheologies and boundary conditions. However, the results of constant velocity boundary conditions are perhaps more generally applicable in plate-driven shear zones, and may give an insight into first order processes in thermal–mechanical shear zones. In Appendix C, we discuss that the constant velocity boundary condition together with the condition for thermally insulating boundaries also is a necessary and sufficient condition for the existence of a homo-
geneous solution. This ensures that the localization problem is well-posed and thus suited to investigate the effect of rheology. The problem is free from localization phenomena caused by a priori chosen geometrical boundary conditions.

Under constant velocity boundary conditions, the linear viscous system self-organizes into a stable ductile shear zone with the following properties. A maximum temperature is achieved inside the thermal–mechanical shear zone whose magnitude does not change with time but varies with rock type. The thermal anomaly broadens conductively leading to a change in width of the shear zone with the square root of time and to an overall weakening of the system. The most striking phenomenon is that, independent of initial rock strength, the viscosity in the center of the shear zone equilibrates to the same minimum. The viscosity minimum in the center of the shear zone is attained when shear heating and conduction are in equilibrium giving the following thermal–mechanical shear zone viscosity (Yuen et al., 1978):

Shear viscosity after feedback \[ \eta_{\text{min}} = 8k \frac{R T_{\text{max}}^2}{Q u_0^2} \] (35)

For a stronger rock with a higher activation energy \( Q \), shear heating is more vigorous so that the maximum temperature in the shear zone \( T_{\text{max}} \) is also higher. This in turn leads, according to Eq. (35), to the thermal–mechanical viscous strength compensation after feedback. Hence, the viscosity of the shear zone is found to be controlled by the initial shear velocity \( u_0 \) of the one-dimensional shear cell. For plate tectonic conditions, this thermal–mechanical feedback viscosity is around \( 5 \times 10^{19} \) Pa s. For a nonlinear power-law rheology, the same phenomenon has been described and the shear zone also converges to a quasi-steady state for which Eq. (35) gives an approximate solution (Fleitout and Froidevaux, 1980). Nonlinear viscous shear zones are narrower and have a more realistic width of \(<20 \text{ km} \).

3.2. Viscous thermal–mechanical and grain size feedback

The analysis has been extended to include an additional feedback between the momentum and rheology. This analysis is based on the observation of extremely fine-grained shear zones in experimentally deformed polycrystalline dunite (Post, 1977). It was already noted that a switch in flow mechanism can lead to strain rate weakening. The feedback is already embedded in Eqs. (16) and (17). A reduced grain size implies a lower flow stress, hence if the first term in Eq. (17) outweighs the second term implying faster grain size reduction than grain growth a condition for mechanical instability is given.

Incorporating this feedback in addition to thermal–mechanical feedback would lead to enhanced focusing of shear zones into a width of several hundred meters (Kameyama et al., 1997). The shear zone is stable under constant velocity boundary conditions. The important factor defocusing the shear zone is the second term in Eq. (17) characterizing grain-growth. It is obvious from Eq. (17) that shear heating implies a grain growth and thereby diminishes the role of the grain size sensitive mechanism. A careful re-investigation of the grain size sensitive mechanism (Braun et al., 1999) shows that the mechanism is probably not a universal shear zone mechanism. In the last section we will discuss, however, a jerky flow scenario where grain size sensitive creep interchanged with thermal heating pulses from ductile earthquakes can keep a shear zone localized.

3.3. Visco-elastic approaches

We now investigate whether, for more realistic rheologies, shear zones are unconditionally stable under constant velocity boundary conditions or whether runaway instabilities are possible. Thermal runaway occurs, if the temperature increase owing to shear heating leads to rheological weakening, feeding back by increasing increments of shear heating. If the feedback runs faster than the conduction can cool the shear zone, an explosive heating phenomenon is possible (Gruntfest, 1963). Thermal runaway has first been suggested as a mechanism for deep earthquakes by Orowan (1960). However, no quantitative proof of the mechanism has been given until Ogawa (1987) investigated a one-dimensional model of a visco-elastic shear zone. He extended the above approach by adding elastic strain rates via the additive strain rate decomposition (Eq. (31)). The addition of elastic deformation implies that the material around the
localized shear zone can act as a storage device for elastic deformation to be released in the ductile shear zone upon instability. Thus until the elastic energy is used up, a situation that is akin to the constant stress boundary conditions, can arise.

Using appropriate parameters for subducting slabs, visco-elastic thermal runaway is possible under constant velocity boundary conditions if the following conditions are fulfilled (Ogawa, 1987): the shear stress must be larger than 300 MPa, the strain rate in the shear zone exceeds $10^{-14}$ s$^{-1}$, the shear zone width is smaller than 100 m and the viscosity inside the shear zone is three orders of magnitudes smaller than outside the shear zone. These conditions are not outside a reasonable geological parameter range for ductile—so called “mylonitic” shear zones (Handy, 1994). However, the required small width of the shear zone has not been resolved numerically. Recall that the power-law rheology used by Ogawa would predict typical quasi-steady state shear zones with a width of the order of a kilometer (Fleitout and Froidevaux, 1980).

Ogawa’s analysis has been extended (Kameyama et al., 1999) to include the Peierls stress-regime (Eq. (19)) in order to investigate whether the addition of this mechanism promotes or stabilizes ductile thermal–mechanical failure. When comparing the temperature sensitivity of the Peierls stress mechanism to that of the power-law, we find that at constant strain rate the stress decreases with the inverse of the square root of increasing temperature in the Peierls stress case (Eq. (19)). This implies a close to linear temperature–stress weakening relation in the Peierls stress mechanism. In the power-law, on the contrary, the exponential Arrhenius term (Eq. (15)) implies an order of magnitude higher weakening for the same temperature increment. In an additive power-law Peierls stress rheology we would therefore expect the Peierls stress to stabilize the thermal–mechanical shear zone while the power-law would be prone to runaway instabilities. This is in fact what has been found out in the Peierls stress analysis (Kameyama et al., 1999). The stabilizing effect of the Peierls mechanism is not strong enough, however, to prevent thermal runaway under constant stress boundary conditions. We will show in the discussion on 2D shear zone models that with the additive strain rate approach ductile thermal–mechanical instabilities are attracted to the transition from Peierls stress-dominated to power-law-dominated creep.

3.4. Elasto-plastic approaches

The shallow counterparts of ductile shear zones are brittle fault zones. Brittle faults typically have a fault width much narrower than their ductile counterpart. Momentum-rheology feedback of non-associated plasticity ensures individual faults narrower than a meter scale, yet brittle faults can extend over several tens of kilometers length. Although brittle fault zones are not triggered by thermal–mechanical feedback, the shear heating term in the energy equation still holds. Furthermore, taking the high speed of a seismic event into account conduction can be neglected. We may speak of “quasi-adiabatic” conditions. It is not surprising at all that brittle seismic events can lead to melting on the fault plane (McKenzie and Brune, 1972).

In the one-dimensional approaches discussed so far, strain hardening has been neglected. The above analyses assume, without saying this explicitly, conditions close to steady state. These are the conditions for which the creep laws have been devised. When looking at Fig. 2, we can see that shear zones most likely nucleate during transient creep in the first bump of the stress–strain diagram and they are fully established under steady state conditions, i.e. the long straight line of the diagram for which the creep laws apply. Only very few analyses have been done with the consideration of transient creep. This is due to the lack of data on strain hardening. The linear stability analysis laid out in the chapter on feedback has been used with the upper bound estimate that the stress after strain hardening cannot exceed Young’s modulus (Hobbs and Ord, 1988). Even under these highly hypothetic conditions, quasi-adiabatic shear zone formation due to thermal–mechanical feedback is not inhibited. We must still verify whether the adiabatic condition represents a good enough approximation and we should follow how the instabilities develop nonlinearly through time. The finite amplitude regime, of course, cannot be determined from a linear stability analysis.

Recently, the potential for adiabatic shear zone formation has been analyzed for geological conditions using the power-law hardening model presented in
Eq. (12) in conjunction with the exponential temperature term of Eq. (15). Again, the equation of momentum conservation (Eq. (6)) and the energy Eq. (14) are used to solve for the dynamic evolution of the shear zone. As a new element a new scaling length $L$ is introduced that is very much larger than the shear zone width (Roberts and Turcotte, 2000). This new scale serves as a description of the integrated elastic area around the shear zone which can store elastic energy. This elastic container adds an external stress to the shear zone, governed by $\mu$ the elastic shear modulus of the elastic container and $w$ the shear displacement at the boundary of the slip zone. Hence, the new force equilibrium is given by:

$$\text{Force equilibrium} \quad \sigma' = \sigma'_0 - \frac{\mu w}{L}$$

where $\sigma'_0$ is the initial shear stress. It follows that under a constant velocity boundary condition a seismic instability nucleates for a critical shear heating (strain rate–energy density) or shear velocity. An astonishing result is that the temperature increase during the seismic instability is not large but only of the order of 20 K (Roberts and Turcotte, 2000). The integrated effect over a geological time scale is similar to the results discussed for Newtonian thermal–mechanical shear zones.

4. Geodynamic modelling (summary of previous work)

Geodynamic models of shear zones traditionally have been separated into two different categories. One group of modelling approaches is guided by observational data of geodynamic processes and the other approach by the physics of the processes underlying geodynamics. The former approach thrives at finding a numerical method that adequately describes a given observation, a top to bottom approach, while the latter investigates fundamental modes of geodynamics from the bottom up. In the former approach, a downscaling scheme is used while the latter uses an up-scaling theorem laid out in Tables 1 and 2.

This paper summarizes the theoretical framework of modeling shear zones, i.e. it focuses on methods that generate shear zones using the up-scaling scheme. We describe briefly the inverse method of manually inserting shear zones from observations. We feel that the basic theory still deserves some more attention before inverting for microphysical parameters from large-scale geodynamical data sets. The future of geodynamic modelling lies without doubt in a solid-/fluid-mechanical approximation augmented by a chemical equation of state for lithosphere and mantle rocks. This needs to be compiled into a unified solid earth reference database (Montesi and Zuber, 2002). Investigation of shear zones from both angles of attack would then be meaningful across different disciplines and not only understood by few experts who are aware of the significance of the basic assumption in the simplified downscaling approaches. We therefore restrict ourselves to reviewing the basic developments in the field without specific application to case studies. We will discuss in the second part only one basic application, which is the quest for self-consistent plate tectonics.

4.1. Viscous modeling

4.1.1. Constitutive theory

Viscous modeling of lithospheric deformation has been introduced to understand the paradox of continental plates that deviate, when they collide, from the basic paradigm of plate rigidity. The first analysis of ductile plate tectonics was by Bird (1978), who pioneered finite element techniques to model the Zagros collision zone, which displays both bulk shortening and a localized shear zone. For a successful model run, Bird found that manual fine-tuning of a low viscosity shear zone was necessary to appropriately model the behavior of the Zagros zone. This fundamental work was followed by investigations of the India Asia collision (England and McKenzie, 1982; Vilotte et al., 1982), where the “soft” Asian lithosphere was modeled by a nonlinear viscous power-law fluid and the Indian indenter by a kinematically driven boundary condition. Since the size of the finite element mesh was very large (several tens of kilometer mesh size or larger) and a fluid rheology was used without feedback no discrete shear zones developed self-consistently. The predicted results yield a smoothed version of the observed deformation. Shear zones have to be implemented manually. Rather than using predefined zones of weaknesses an apparently
Similarly, shear zones were not found as a natural outcome of semi-analytical calculation of sinusoidal stretching instabilities of a non-linear fluid lithosphere (Fletcher and Hallet, 1983; Ricard and Froidevaux, 1986). Yet, shear band formation growing out of such stretching instabilities is a classical experiment in material sciences. Ductile shear zones are known to form in materials that localize less readily than rocks, e.g., metals. However, for numerical modeling of such shear zones, some form of softening has to be considered (Needleman and Tvergaard, 1992). A fluid rheology without a yield like phenomenon or feedback does not have the potential for forming shear zones. At best, a fluid may be used as an up-scaled version of discontinuous deformation with accepting the uncertainty that some basic physics is missing. While in the early 1980s, this approach was logical because computational power was limited; nowadays, there is no reason to use fluid approaches for the lithosphere. Exceptions are cases where coupling to convection in the mantle poses a large additional computational workload (discussed below), or where a smoothing of boundary conditions is the desired effect (Marotta et al., 2001), or in laboratory experiments where similarity criteria restrict the availability of analogue experimental materials (Faccenna et al., 2001).

### 4.2. 2D visco-plastic modeling

#### 4.2.1. Constitutive theory

Including a plastic limit stress by a pseudo-plastic or a Bingham style formulation alone does not turn a fluid model into a simulation with a shear zone developing. Although strain hardening is zero, it is not negative and the closest equivalence of a shear zone occurs when strongly heterogeneous boundary conditions lead to strongly heterogeneous shear flow with some areas that hardly deform and others that deform vigorously. This was in fact what was found out in early calculations with a co-axial visco-plastic formulation (Bird, 1988). When modifying the yield phenomenon into a non-coaxial one, dramatic changes are observed (Beaumont et al., 1996a; Ellis et al., 1999; Lenardic et al., 2000).

Shear zones form readily in such a medium, the only drawback being that non-associated plasticity (Appendix B) applies to brittle fault zones only. Because of the weak strength and the small thickness of the brittle layer (Fig. 7), it is not an appropriate model for the entire lithosphere. Another drawback of this approach is that shear zones are rather fickle features. New shear zones form readily during deformation and old shear zones disappear altogether. A way out of this problem is to self-lubricate shear zones by parametrically imposed strain rate weakening laws (Bercovici, 1993) or strain softening laws (Govers and Wortel, 1995; Huismans and Beaumont, 2002). Both formulations preserve shear zones once they are formed. The advantage of the strain weakening over the strain rate weakening law is that in the former approach shear zones will retain their memory long after deformation ceases, while in the strain rate weakening case shear zones are instantaneous features and vanish after a change of boundary conditions (Zhong et al., 1998). One could argue that the parametric weakening curve records the history of feedback without any explicit modeling of feedback. However, this would require a re-parameterization of fully coupled feedback calculation, which has not been done, to date.

#### 4.2.2. Energy theory

Fluid dynamic calculations of visco-plastic shear zones with coupled feedback phenomena have been performed with varying degree of dynamic self-consistency. Faults have been added as a frictional boundary constraint to a viscous mantle (Zhong et al., 1998) and the degree of shear heating close to the boundary has been recorded in a kinematic model (van Hunen et al., 2000). It turns out that shear heating can accelerate deformation by about 20% through lowering the viscosity down to $10^{20}$ Pa s. No dynamic heating event was recorded in these predefined fault zone models. A more significant effect has been found for calculations in which a spasmodic release of significant gravitational potential energy into shear heating was possible (Schott et al., 2000). The shear zone was left unconstrained. However, none of these models was able to predict clear shear zones out of the fluid dynamic approach. The image conveyed by these calculations is that of highly transitional features, which makes these approaches a poor candidate for modeling plate tectonics. When considering void-volatile feedback, the picture changes (Bercovici, 1993).
Stable narrow shear zones generate self-consistently from pure volatile feedback.

4.3. 2D rigid-plastic modeling

4.3.1. Constitutive theory

Prior to the use of numerical models in geodynamics, a simple analytical technique the so-called “slip line field method” has made a brief appearance in geodynamic modeling of shear zones (Tapponnier and Molnar, 1976). It is practically not used at present apart from sporadic reappearance in the literature (e.g. Regenauer-Lieb and Petit, 1997). The slip line field method is only available for co-axial deformation with simple plane strain or plane stress boundary conditions. The rheology is simplified to a simple rigid-plastic body. A static plastic equilibrium must exist, i.e. the method cannot be extended to dynamic equilibrium. Finally, being (semi-) analytical the solution to problems of complex geometric boundary conditions is tedious. Yet, prior to the availability of numerical methods, the theory has governed 20 years of metal deformation and has led to the rapid advance of the theory of plasticity (Johnson et al., 1982).

Slip lines are the characteristics of hyperbolic differential equations (see Appendix C). They are the perfect mathematical embodiments of ductile shear zones since they are capable of predicting either continuous shear or discrete shear velocity discontinuities, i.e. vanishingly thin shear zones. What makes the method so invaluable is that it is the only analytical method that allows the prediction of shear zones in two dimensions. For the plane stress case, displacement can be three-dimensional. Since the existence of velocity discontinuities simplifies the plastic solution technique, the strength of the method relies on the weakness of the numerical methods. In the engineering community the method is therefore routinely used for benchmarking new numerical techniques (Li and Liu, 2000).

Although the method does not, by definition, allow for feedback it can be used to predict potential lines of ductile failure that may or may not develop in a ductile body with a less ideal rheology. Examples are slip lines appearing as lines of dilatant fracture (Coffin and Rogers, 1967), as lines of martensitic transformation (Rogers, 1979) or slip lines appearing as heat lines through shear heating feedback (Johnson et al., 1964). The heat line approach has been used to predict a time averaged maximum amount of shear heating throughout the last 10 million years of collision in the Himalayas (Hochstein and Regenauer-Lieb, 1998). The collisional energy dissipated on heat lines was found to be sufficient to maintain the observed anomalous heat transfer in convective geothermal systems in the Himalayas. A large portion of this energy appears to have been stored during initial mountain building processes as gravitational potential energy now released by extension (Tanimoto and Okamoto, 2000).

4.4. 2D elasto-plastic modeling

4.4.1. Constitutive theory

In the case with elasticity, the tendency for localized shear zones, so prominent in the rigid-plastic approach, is wiped out significantly. This is due to replacing infinitely thin rigid-plastic shear zones by zones of finite width of an elasto-plastic material. Although this analysis lends itself perfectly for the investigation of realistic ductile shear zones, elasto-plastic modeling of lithosphere deformation has traditionally been carried out without any focus on shear zone development. Exception are analogue laboratory studies (Shemenda and Grocholsky, 1992) or studies of fault zones in the brittle field. An excellent review of the studies in the brittle field can be found in Gerbault et al. (1998). The prime interest of elasto-plastic studies including the ductile field has been focused on assessing flexural rigidity (see Albert et al., 2000). Therefore, strain hardening/weakening and feedback have usually been neglected. We will discuss strain hardening and ductile elasto-plastic shear zones in the section on applications.

4.5. 2D elasto-visco-plastic modeling

4.5.1. Constitutive theory

The lack of elasto-plastic studies on ductile shear zones in the semi-brittle regime of Fig. 7 has been rendered into the following simplified working hypothesis (Burov et al., 2001; Cloetingh et al., 1999; Gerbault et al., 1998; Moresi and Solomatov, 1998).

1. Ductile shear zone nucleation processes are assumed to be of second order importance.
2. The propensity of shear zone nucleation in the brittle field renders oblivious all other feedback processes.
3. The presence of ductile shear zones indicates a large accumulation of strain within the ductile layer leading to the development of ductile shear zones.
Hence, all shear zones nucleate in the brittle field. Shear zones propagate from the brittle field into the ductile field and peter out. The most significant drawback of this method is that the mechanical role of the brittle layer is overemphasized. In Fig. 7, we can see that the brittle zone is weak and presumably does not extend deeper than about 10 km depth. Thermal–mechanical feedback processes occurring in the strong semi-brittle regime of the lithosphere are completely neglected in these studies.

4.5.2. Intrinsic length scales and the energy theory

Elasto-visco-plastic modelling with feedback includes all ingredients necessary for the investigation of the transient creep phenomenon leading to the nucleation of shear zones. The drawback of this approach is that a proper implementation of the multiphysics requires a large wealth of material data and is computationally demanding (Tables 1 and 2).

The computational cost relies on the high degree of spatial and temporal resolution that is required for resolving the multiscale nature of the feedback. The spatial scale that needs to be resolved before shear zones are visible can be derived from one-dimensional calculations and from theoretical considerations. In the following compilation, we will see that geological observations are much better suited to identify effects of the different length scales than engineering applications. Through geological observation on shear zones, we are in the unique position to pay back some physical insight into the knowledge base compiled in more than 50 years of theory of plasticity. Up to now, the basic progress has been made in metallurgy. Unfortunately, for metals, the intrinsic material length scales of plasticity and thermal feedback (Lemonds and Needleman, 1986) collapses into the micron-scale, while in geology thermal feedback and mesoscale plasticity spreads out owing to the slow deformation and the low diffusivity of rocks. On the issue regarding the nucleation of shear zones, a separation of the length scales for shear zone formation is a fundamental issue.

The intrinsic material length scale of deformation by dislocations can be shown to govern the width of shear bands in metals (see Aifantis, 1987 for a review). The fundamental physics of this length scale hinges on a breakdown of the classical continuum mechanics where dislocation can be referred to as “statistically stored dislocations” while below 10 μm the discrete nature of dislocations is felt and there appear so called “geometrically necessary dislocations” which are related to the gradients of plastic strain in a material. Recently, nano-indentation and micro-torsion experiments have given support to this theoretically postulated limit (Bulatov et al., 1998). It was found that it is 200–300% harder to indent at nanoscale than at large-scale (see Gao et al., 1999 for a review). The immediate outcome of this is that, in plasticity, there appears an intrinsic material length parameter \( l_1 \) characterizing the energy of defects. This defect energy governs the strain gradient of plasticity at mesoscale.

Material length scale of plasticity \( l_1 = M(\mu e^n)^2 b \)  

(37)

where \( M \) is a material parameter (around 18 for metals), \( \mu \) the elastic shear modulus and \( e^n \) is a reference stress coming in from the power-law hardening law of Eq. (12) and \( b \) the Burgers vector which is of nanometer scale. The strain-gradient plasticity theory recovers at large-scale the power-law hardening relationship when a macroscopic population of statistically distributed dislocations is achieved. While this length scale relies on the shear gradients, it has been suggested to expand the theory to include a second length scale for stretch gradients, which would govern a critical void size before void-void coalescence (Fleck and Hutchinson, 2001). All of these length scales are below tens of micrometer scale.

Future analyses of shear bands should include strain gradient effects mapping microscale dislocation interactions into mesoscale cells (Guo et al., 2001). An upsampling of these results then would allow to implement shear bands into large-scale shear zone models. However, a clear separation of thermal feedback (Lemonds and Needleman, 1986) and strain gradient effects (Aifantis, 1987) has not yet been developed.

In geological applications, such a separation is possible. The length scale \( l_1 \) for rocks and ceramics is also of micrometer scale but the thermal length scale is much larger than in metals. In Table 2, we have defined the thermal length scale to be:

Thermal diffusion length scale \( l_2 = \sqrt{\frac{\kappa}{\delta}} \)  

(38)
This length scale is a new aspect of the energy theory. It is the fundamental quantity controlling the final postlocalization equilibration width of shear heating controlled shear zones (Regenauer-Lieb and Yuen, 2003; Sherif and Shawki, 1992) when heat conduction and shear heating are in thermal–mechanical equilibrium. It is also the quantity that governs the resolution criteria for numerical thermal–mechanical modeling of shear zones (Regenauer-Lieb and Yuen, 2003). In order to be able to see thermal feedback in a numerical simulation, we need to resolve below the thermal length scale. Taking e.g. a thermal diffusivity of rocks $\kappa$ to be of the order $1 \times 10^{-6}$ m$^2$ s$^{-1}$ and a strain rate in the shear zone of the order of $1 \times 10^{-12}$ s$^{-1}$, we would obtain a thermal length scale of the order of 1 km. This resolution is achievable in any 2D simulation, even on a plate tectonic scale. The one-dimensional models, discussed earlier, predict for the case of a power-law fluid a thermally triggered shear zone width of initially 2 km width widening with the square root of time (Fleitout and Froidevaux, 1980). We conclude that a 2D numerical approach in lithosphere dynamics needs to have a spatial resolution of at least a kilometer, preferably smaller. The spatial scale that is introduced by diffusion creep is the solid-state chemical diffusional length scale, which depends on the relative size of anions, such as silicates.

Chemical diffusion length scale $l_3 = \sqrt{D_{\text{eff}}t}$ (39)

where $D_{\text{eff}}$ is the effective diffusivity at a given pressure/temperature and $t$ is the time. When this length scale becomes important, diffusion accommodated creep can become prevalent over deformation assisted by dislocations.

This length scale critically influences the potential for grain size feedback, hence is also an intrinsic quantity of the energy theory of localization. In order to resolve all the physics introduced by this feedback, a numerical simulation would have to reach the scale of the minimum grain size in the system upon which diffusion can operate. Referring to Table 2, an upscaling formalism has been suggested that describes the flow of a statistical population of grains by a continuum with a linear viscous flow law. In the calculations of Kameyama et al. (1997), a spatial resolution of 1 m has been reached. The predicted shear zone due to feedback is in this case on the order of hundreds of meters. Assuming that the smaller scale physics does not change the behavior of the system, we conclude that a 2D numerical approach needs to have a spatial resolution of at least a hundred meters.

The spatial scale introduced by the void-volatile feedback is on the order of a fluid inclusion (say 50 $\mu$m). This length scale can be introduced into strain-gradient plasticity through consideration of an additional length scale from stretching strain gradients. However, such a resolution is beyond reach for geodynamic calculations but may be linked by discrete particle method, such as smoothed particle hydrodynamics (Monaghan, 1992).

Again, an upscaling scheme has to be used. Here we assume normal void volume populations through the population density parameters $A$ and $B$ in Eqs. (26) and (27) for the ductile and brittle void nucleation cases. Furthermore, in treating void volume as a smeared continuum within a particular finite element, any smaller scale physics is suppressed. Using finite elements with a size of $200 \times 200$ m, the void-volatile feedback predicts relatively wide void sheets driving a fluid-filled geodynamic shear zone with a width of about 10 km (Regenauer-Lieb and Yuen, 2000b). Ignoring possible feedback mechanisms at smaller scale, we recommend a minimum resolution of 10 km for a void-volatile feedback calculation.

In sum, we would want to have a maximum element size of the order of 100 m in order to be able to resolve all feedback mechanism within a single numerical analysis. Now, a typical 2D geodynamic calculation would comprise an area of $1000 \times 100$ km. This would imply about 10 million nodes in the calculation. It becomes apparent why ductile shear zones are hard to capture in geodynamical calculations. Ductile shear zones are however not beyond the reach of current computers.

4.5.3. Energy theory

The toughest candidate for 2D shear zones is undoubtedly the grain size sensitive feedback. The only 2D calculation with grain size sensitive feedback done so far focused on the physics of grain size sensitive creep in convection. Therefore, it had a local resolution of only about 5 km (Hall and Parmentier, 2003). This scale exceeded the spatial resolution requirement for shear zone nucleation by an order of magnitude. Consequently, in contrast to the 1D cal-
culations, no zones of highly localized deformation were observed for obvious reasons.

Although the physics appeared to be fully implemented, the first 2D calculations with thermal feedback (Chery et al., 1991) missed entirely the phenomenon of thermal–mechanical shear zones. The size of the finite element discretization was chosen larger than $l_2$. In order to be able to resolve high strain rates and avoid undesirable mesh effects, at least four elements should be contained within $l_2$. The implied 200 m resolution was implemented in an idealized 2D setup of a perfectly homogenous isothermal elasto-visco-plastic olivine plate under constant extensional velocity boundary conditions (Regenauer-Lieb and Yuen, 1998). Analogous to the one-dimensional calculation a local perturbation in the form of a weak inclusion was used (Regenauer-Lieb and Yuen, 2000b). The schematic layout and the predicted sinistral and dextral shear zones are shown in Fig. 8.

We can now compare this elasto-visco-plastic 2D calculation with the 1D elasto-plastic calculation of Roberts and Turcotte (2000). Our formulation is in fact the visco-elasto-plastic equivalent of the 1D calculation. The elastic scaling length $L$, which represented an elastic container around the shear zone in the 1D calculation, is implemented explicitly in the 2D calculations. It stores elastic energy during the charge up time of 600,000 years during which about 12 km elastic stretching of a 1000 km long elastic layer occurred. Finally, the plastic threshold stress is reached near the imperfection and the stored elastic energy is released in seismic failure of the lithosphere. It is surprising that only a moderate temperature rise of a few tens of degrees is necessary to cause ductile failure of the olivine sheet. Another important aspect is that thermal runaway leading to melting instabilities is not expected. In Figs. 9 and 10, we have plotted the results of the thermal–mechanical calculations of Regenauer-Lieb and Yuen (2000b) and Roberts and Turcotte (2000), respectively, showing the full evolution of the seismic event. It is clear that the seismic event terminates before 25 K shear heating have been achieved. The addition of elasticity therefore has led to a dy-

![Propagating shear zones after 0.8 Ma extension](image)

Fig. 8. Isothermal Olivine Plate under constant plane strain extension. The olivine plate has dimensions 1000 × 100 km. Analogous to the one-dimensional models a small weak imperfection was introduced as a nucleation point for shear zones. Only thermal mechanical feedback of power law flow law and void volatile feedback were considered. The shear zone propagates rapidly through the plate after storing elastic strain energy for 600,000 years. The same time lag has been reported in one-dimensional models of ductile seismic instabilities in metals (Shawki, 1994a,b). The void-volatile damage zone is trailing the thermal-mechanic feedback mimicking its crack like shape. After 800,000 years, the shear zone turns into a seismic event shown in Fig. 9.
namic instability, which after some time reaches the same order of magnitude of shear heating as the purely viscous shear zones.

In summary, we can infer that during the lifetime of shear zones several feedback mechanisms play a different role at different times. The first feedback mechanism is the momentum-rheology feedback. An elastic wave has been monitored (Regenauer-Lieb and Yuen, 2000b) to propagate ahead of the thermal–mechanical plastic wave which is shown in Fig. 8. Depending on the energy stored in the elastic envelope, the second thermal–mechanical plastic wave turns seismic or not. Structural damage follows in both cases mimicking a thermal–mechanical Mode II crack-like feature. Structural damage occurs either through void-volatile interaction or grain size reduction, thus engraving the shear zone for larger time scales. In considering the multiphysics of shear zone formation and their demand for spatial resolution (below 100 m) and temporal evolution (below 1 s) it becomes clear that 3D calculations are not yet ripe for any sensible undertaking, unless adaptive wavelet methods are employed (Vasilyev et al., 2001).

4.6. 3D modeling

Three-dimensional calculations do not belong to the classes of modeling discussed so far because they attempt at solving a particular geodynamic problem without going systematically through the physics of the processes underlying shear zone formation. The extreme spatial and temporal resolution demand posed by the feedback calculations is circumvented by manually introduced shear zones or by postulating simple parametric rheological models or by considering only numerically tractable feedback mechanisms. Three-dimensional approaches to the fundamental problem of self-consistent plate tectonics from mantle convection calculation are good examples (e.g. Tackley, 1998; Trompert and Hansen, 1998) and shall be discussed below. Other approaches focus on thermal–mechanical feedback within the convecting mantle (e.g. Bala-
chandar et al., 1995; Dubuffet et al., 2000) or attempt to solve the problem of nucleation of shear zones in an intraplate volcanic field by void-volatile interaction alone (Regenauer-Lieb, 1999). A pioneering three-dimensional shear zone model of the San Andreas fault zone has been dealt with by Williams and Richardson (1991) using visco-elastic rheology while 3D modeling of the plate-mantle interaction problem has been first tackled with application to the Australia-Antarctica subduction system (Gurnis et al., 1998).

For practical geodynamic purposes, simplified approaches need to be developed. The key questions that need to be addressed are: Is it possible to neglect elasticity to suppress the tendency of the fully coupled system to turn into a seismic instability, can strain hardening be neglected, can plasticity be neglected? In the following attempt at solving the plate tectonic coupling problem, we will go through the different approaches and show the importance of the individual rheological ingredients.

5. Geodynamic modeling applications

Understanding plate tectonic coupling has been a core question addressed in the geodynamic community in the past 10 years. At the heart of this problem lies the observation that plate boundaries are the largest shear zones on the Earth (Gordon and Stein, 1992). They can last for hundreds of million years and if they ever get inactive for some time they can be reactivated at a later stage. What then causes the nucleation of new plate boundaries? How can old plate boundaries be reactivated? How can a plate boundary survive for an extended geological time period? How is the plate like motion coupled to convection in the mantle?

The following discussion does not aim at giving a review of the generation of plate tectonics from mantle convection, but uses the ongoing discussion summarized in review papers (e.g. Bercovici, 2002) as a way to illustrate the fundamental limits of the fluid rheological approaches that have been proposed. We then focus on the basic problem of subduction initiation for which solid mechanical models are available. We use this as a common platform to discuss the central role of elasticity, shear heating and water for generating lithosphere scale faulting.

5.1. Visco-plastic plate tectonics

When looking at the long time scale of plate tectonic cycles, it appears at first sight legitimate to neglect elastic strains and only consider the role of viscosity and plasticity. Implementing plasticity into standard viscous mantle convection calculations hence has been the main stream of attack. An example for a 3D fluid-dynamic calculation (Trampert and Hansen, 1998) that reproduced plate-like behavior of the top surface by considering a Bingham-type rheology is shown in Fig. 11.

The basic deficiencies of the model are immediately clear. The plate boundary is still diffused, i.e. no discrete shear zone develops, there is only very little vertical axis rotation (toroidal flow component is too low), the strength of the lithosphere is too low, the downwelling is two sided and the system does not keep a permanent lithospheric identity. From time to time, convective instabilities drag the entire lithosphere-like material into the mantle. The calculations seem therefore more apt at describing a scenario that has been postulated as a resurfacing event on Venus (Grosfils and Head, 1996).

To improve some of these deficiencies, a systematic analysis of the yield stress has been performed (Tackley, 2000a,b). In these calculations, the lithospheric yield envelope pictured in Fig. 7 was parameterized in a pseudo-plastic flow law. The pseudo-plastic formulation rather than the Bingham visco-plastic body allows increasing of the yield stress without going suddenly into the stagnant lid regime. Recall that the yield stress in the Bingham body (Fig. 4) acts as a toggle switch between zero deformation below and sudden deformation above the yield stress. The pseudo-plastic law on the contrary allows some very small deformation before the yield stress is reached (Fig. 5). This minute detail is very important and allows a range of coupling between mantle convection and lid that cannot be observed in the Bingham approach. It was found that the brittle strength contributed little to the overall behavior of the lithosphere. Plate-like results were achieved by a constant strength in the ductile part of the lithosphere. If partial melting and associated low viscosity asthenosphere allows for additional decoupling of this stiff layer, a plate tectonic scenario can be obtained self-consistently (Fig. 12).
While this approach recovers more of an Earth-like dynamics than the approach shown in Fig. 11, some important shortcomings still remain. The yield stress is higher but still too low when compared to laboratory analyses (Fig. 7). Also, the deficiency that subduction is near vertical and has double sided downwelling could not be resolved. Finally, no pure strike slip faults exist. What these calculations clearly show, however, is the importance of the yield stress of the lithosphere. In the following we will focus on the question how to destroy the integrity of the lithosphere and form a new plate boundary. Since spreading centers seem to be well resolved by the above visco-plastic calculations, we will home in on the problem of subduction initiation as a key player in Earth dynamics. Ultimately, we would also want to abandon parametric approaches and merge them with more complete rheological results from the laboratory.

5.2. Elasto-plastic passive margin evolution

As a first step we use a parametric power-law elasto-plastic hardening model (Eq. (12)) and systematically vary the yield stress and the strain-hardening power-law exponent. Elasticity is considered by coupling elastic and plastic deformation in the Ramberg–Osgood approximation (Branlund et al., 2001). The Ramberg–Osgood approximation is a non-linear elastic fracture mechanical approach that does not separate plastic from elastic strain. The results are compared to the additive strain-approximation (Eq. (31)). Following an earlier suggestion we investigate whether sediments loaded onto a passive ocean continent boundary (OCB) can cause failure of the lithosphere (Cloetingh et al., 1982).

We have already discussed strain hardening in the chapter on length scales. Strain hardening is a fundamental property of continuum mechanics communicating microscopic discontinuous deformation at nano-scale into a macroscopic plastic flow law. As plastic strain increases, so does the dislocation density. This leads to dislocation interaction, which in turn is influenced by dislocation mobility. Metals can be shown to have a power-law exponent that lies between 3 and 7 (Hirsch, 1975). Rocks and ceramics have silicate covalent binding, which are difficult to break. They have a high tendency for micro-brittle failure at low temperatures. At higher temperature it is easier to break the binding and the deformation by dislocations increases. The importance of the strain-dependent dislocation state in the temperature range 500–800 °C, has been neglected in the geological literature. However, pioneering work by Griggs et al. (1960) shows that strain hardening is very small for
olivine. An adequate fit of the experimental results has been obtained with a relatively high $n$ of 35 (Branlund et al., 2001).

The different hardening laws have been applied to the passive margin model of Fig. 13 and a snapshot after 60 Ma loading is shown in Fig. 14. The strain-hardening exponent controls the stiffness of the plastic response with a higher stiffness for higher $n$. This analysis gives some insights into the effects of strain hardening and it also shows that more realistic stress levels for the wholesale failure of the lithosphere can be achieved through dynamic interaction of elasticity and plasticity. Although the elasto-plastic model gets closer to the laboratory strength curves, it does not quite reach the laboratory strength estimates. The obvious solution to this problem is to consider viscous deformation in the bottom part of the lithosphere.

5.3. Elasto-visco-plastic passive margin evolution

The idea of visco-elastic stress amplification as a means to clip the high strength branch of the yield stress envelope has been promulgated by Kusznir (1982). The physics underlying visco-elastic stress amplification is simple. Because the lower part of the lithosphere can flow more readily than the upper part it will, under an applied stress, deform by viscous deformation. This consequently increases the elastic stress field immediately above the flowing portion until the yield stress is reached. Upon repeating this process in the higher levels, the high strength branch can be continually eroded and failure of the entire lithosphere appears possible. The idea has been tested for the case of passive margin evolution (Branlund et al., 2001; Regenauer-Lieb et al., 2001).

Fig. 12. Plate tectonic simulations with a constant yield stress pseudo-plastic lithosphere (Tackley, 2000b). The yield stress is indicated on the temperature iso-surface plot (right column). The left column shows a viscosity plot with the color bar ranging from purple (lowest viscosity) to red (highest viscosity). The system goes from distributed divergence with localized downwelling at low yield stress (34–70 MPa) to sharper spreading centers and optimum Earth-like toroidal flow at higher yield stress (103–150 MPa) until at 220 MPa the system switches to an episodic rigid lid regime and finally at 340 MPa mantle convection is covered by a stagnant lid.

Fig. 13. Model setup to test the influence of strain hardening on ductile failure of passive margins (Branlund et al., 2000). The lithosphere is loaded incrementally by an increasing sediment load with a peak of 15 km after 100 Ma loading. The elasto-plastic lithosphere is supported by a quasi-elastic foundation where the spring stiffness is reproducing the buoyancy contrast produced by displacing mantle material through sediments and water.
To illustrate this concept, the base model in Fig. 13 has been modified to incorporate the following effects. The temperature profile of a cooling half space model was added, the lithosphere has a composite visco-elastic dry olivine rheology with diffusion, Peierls and power-law creep incorporated by the additive strain rate decomposition. The brittle top part of 10 km is not considered, i.e. the sedimentary load is immediately imposed onto the ductile portion of the lithosphere. It was found that a fully coupled calculation localizes readily on any heterogeneity in the field. This enhances the prospects for numerical grid artifacts. Therefore the singular peak of the sedimentary loading function was smoothed and adapted to the Western Atlantic passive margin. All nodal loads were replaced by surface loads and the asthenosphere was added as a viscous foundation. In order to investigate whether asymmetric collapse is possible any source of asymmetry other than the oceanic temperature profile and the asymmetric loading function were removed. Hence the geometrical heterogeneity at the OCB was removed. A zoom-in on the highly deforming part of the model is shown in Fig. 15.

Although visco-elastic stress amplification seems to work it causes unexpected decoupled fluid- and solid-like deformation, each with its own intrinsic time-scale. Hence, it does not rupture the integrity of the lithosphere. It is apparent that the lower part of the lithosphere is too weak to deform as a solid entity together with the upper part. Therefore, we obtain a result, where the yield stress appears to be realistic but for subduction initiation we still need to synchronize the fluid and solid deformation in the lithosphere.

5.4. Add water

At this juncture, we may reconsider the two visco-plastic models introduced in Figs. 11 and 12. The model by Trompert and Hansen considers a Bingham visco-plastic body and the model by Tackley a pseudo-plastic flow law. Both models reach the stagnant lid regime at extremely different values of yield stress. In the Tackley model it was found that the “best” results were obtained, if the lower part of the lithosphere had a constant yield stress thereby shielding the lower part of the lithosphere from instabilities that are shown in Fig. 15. By combining the Bingham plastic model together with the pseudoplastic model, we can expect that the lower yield stress of the Bingham model shields the lithosphere from recycling at low stress while the upper yield stress reaches the excessive strength expected for a linear Bingham viscous flow law at high strain rates.

Such a combined composite rheology is in fact what is obtained by adding power-law and Peierls stress mechanisms. The lower Bingham style yield stress is embedded in the Peierls stress law (see Fig. 4) and the high stress ceiling is part of the power-law flow (see Fig. 5). Why are we not feeling the shielding effect of the Peierls stress Bingham-like part in our model calculations in Fig. 15?

The key in the success lies in the synchronization of solid and fluid deformation. Solid mechanical
Deformation in shear zones occurs at much faster strain rates than the fluid style deformation in the convecting mantle. The communication between shear zones at surface and the fluid-like deformation at depth must be well coordinated to prevent a temporal lag between deformation at surface and at depth as seen in Fig. 15. If we increase the critical strain rate for the onset of Peierls creep and thus shield the fluid-like layer of the lithosphere then fluid and solid deformation may perhaps be coupled. The principal parameter controlling this strain rate is the water content in the mantle (Eq. (21)). The same model calculation showed in Fig. 15 has been repeated by raising the water content in the lithosphere (Fig. 16). We can see from Fig. 16 that this logic holds. Just by adding water, the shear zone can propagate through the lithosphere instead of curving back to the surface. Therefore, a new type of tectonics appears where Rayleigh-Taylor-like instabilities at depth are shaped by asymmetric shear zone in the top, which are, in turn, fed by the gravitational potential energy release of the negatively buoyant system. Coming back to the feedback diagram in Fig. 7, there is a clear evidence that water regulates the feedback between fluid-like deformation at depth and solid-like deformation at the surface. Water content in the lithosphere and in the adjacent mantle dictates whether or not a lithosphere scale fault can develop. It thereby regulates the style of tectonics in a terrestrial planetary system (Regenauer-Lieb and Kohl, 2003).

6. Viscosity and lifetime of shear zones

A worldwide compilation of plate boundaries and shear zones permits an inverse approach, allowing for
the prediction of the long-term geodynamic strength of plate boundary shear zones from observed plate velocities. However, since shear zones have to be implemented manually into the numerical approach, they have been put in either through idealized velocity discontinuities using a contact friction law on the fault surface or through finite shear zones of lower effective viscosity. The former approach has proven to be more successful (Bird, 1998). A very low value of friction of 0.03 was found on plate boundary shear zones supporting the idea that plate boundaries are indeed weak. More detailed regional analyses of the Africa-Europe plate boundary along the Gibraltar–Azores segment have given somewhat larger friction values of the plate boundary (0.1–0.15) but it still appears to be four times weaker than the adjacent lithosphere (Jimenez-Munt et al., 2001).

While the mathematical idealization of a shear zone by a velocity discontinuity with a sliding friction law is a crude approximation the principal result of weak shear zones cannot be disputed. We investigate here whether the ductile shear zone in Fig. 16 becomes sufficiently weak. For this, we plot a viscosity profile across the middle section of the major left lateral shear zone in Fig. 16 (Fig. 17).

The shear zone is weaker than the model asthenosphere and is indeed weak enough to cause initiation of asymmetric subduction. We conclude that consideration of complete elasto-visco-plastic rheology with thermal feedback resolves all of the deficiencies reported in the above self-consistent approaches to plate tectonics. We emphasize, however, that the addition of water is just as important as thermal feedback (Regenauer-Lieb and Kohl, 2003). Unfortunately, for reasons of excessive numerical cost, these high resolution calculations can at present only be done in 2D. Surprisingly, the viscosity inside the shear zone is of the same order of magnitude as predicted by the simple one-dimensional viscous feedback calculations (Eq. (35)). This raises hopes of parameterizing a simpler 3D approach with high-enough resolution, which can be benchmarked by a complete 2D calculation.

Next, we discuss the lifetime of shear zones. We have recognized the importance of shear heating in the nucleation phase of shear zones. We have also seen

Fig. 16. Same as Fig. 15 but water has been added (COH = 810 ppm H/Si) (Regenauer-Lieb et al., 2001). Void-volatile and grain size sensitive feedback are not considered. Solid and fluid deformations are coupled, and the lithosphere fails on its entire thickness. Ductile fault zones develop dynamically. The major sinistral shear zone rotates counterclockwise during its evolution. A subsidiary sinistral fault is developing to the left of the first hinge like shear zone. The top of the plate is weakened by zigzagging shear zones.
that shortly after the nucleation phase structural damage will swamp the thermal–mechanical feature (Fig. 8). It is obvious that structural modifications are guaranteeing the longevity of fault zones and their potential for reactivation. When weighing in the two structural mechanisms, i.e. void-volatile versus grain size sensitivity, it is obvious that the void-volatile mechanism is better suited for creating shear zones on geological time scales. The diffusion of volatiles out of the shear zones is very much smaller than the healing through grain growth in grain size sensitive creep or spreading of the anomaly through thermal diffusion. This applies to carbon dioxide but not to water because of the abundance of hydrogen related point defects (Kohlstedt and Mackwell, 1998). The bulky molecules of carbon dioxide in contrast will remain trapped within the shear zones (Nakazaki et al., 1995). This explains the observation of abundant CO₂ inclusions in xenoliths (Roedder, 1981). Hence, we suggest to use water content as a global variable and within shear zones consider only the void–void interaction formulated in the section on volatiles.

Sheets with preferentially aligned CO₂ voids in the mantle are not the only factors that could guarantee the longevity of shear zones. Other structural modifications have been suggested. Structural heterogeneity of the continental lithosphere (Tommasi et al., 1995) or the mechanical anisotropy of olivine within the mantle part of the lithosphere (Tommasi and Vauchez, 2001) have been shown to preserve shear zone memory and cause nucleation of shear zones in preferred orientations. The longevity of shear zones through macro-scale geological and mesoscale structural heterogeneity is a natural result of structural geological observations.

They form a step up in scale of the discontinuous processes discussed so far. Unfortunately no rigorous formulation has been developed. There have been first attempts at describing heterogeneous steady state creep through their energetics. Consider, for instance, a two-phase strength system, a dynamic evolution towards an interconnected layer of the weak crystals (Handy, 1994) during shearing, can embed a weak fault into a structurally more competent host rock. The dynamic evolution process of shear zone nucleation-growth and coalescence of weak phases is formally equivalent to the mathematical concepts of two-phase flow introduced in the void-volatile mechanism (Bercovici et al., 2001a). We conclude that heterogeneity is a prime candidate to produce and preserve ductile shear zones on geological time scales. Macro-scale structural heterogeneity evolves dynamically and draws on the non-thermal energy fraction \( \frac{1}{C_0 v} \) of the deformational work stored inside the thermal–mechanical shear zone.

The concept of heterogeneity brings us now to the brittle field. Being potentially more heterogeneous than the ductile part of the lithosphere, it is necessary to look into the equivalent local approaches to fracture also known as “damage mechanics”. Damage is stored as an additional internal variable and considers the dynamic evolution of structural heterogeneity. We have already looked into a ductile class of damage mechanics, which we recommended to calculate semi-brittle and ductile faults. In this case, the damage state variable is the void–volume ratio (Eq. (22)). Brittle faults can also promise longevity and an equivalent state variable has been introduced that describes a population of brittle microcracks. A complete damage mechanical model for rocks has been developed by Lyakhovsky and co-workers (Lyakhovsky et al., 1993, 1997). The advantage of damage mechanics over the classical fracture mechanical approach is that it is
amenable to prolonged histories of brittle crustal evolution with a complicated dynamic interaction between local damaged zones, which are modeled by zones of degraded elasticity. The disadvantages are additional difficulties to formulate an objective energy flow rate into the cracks and the problem of mesh sensitivity. This approach, also dubbed “smeared crack” approach, has reached a high level of sophistication in applications based on concrete mechanics (de Borst, 2002). Following this line of attack into an Earth-like scenario is a promising field for understanding brittle fault zone dynamics.

Hybrid models that combine this method together with discrete elements, modeling explicit cracks from classical fracture mechanics, have been formulated giving realistic fracture patterns. The success lies in using discrete element and finite element methods together, because the discrete element method considers discontinuous deformation and the finite element model stores the continuum. This hybrid global–local model unfortunately comes up with excessive computation demands only to be realized in grand challenge computations with topline computers since 1 mm resolution has to be achieved to avoid mesh size effects (Munjiza and John, 2002). The same applies to particle codes discussed below.

The longevity and memory of fault zones remain an unsolved problem in plate tectonics. While structural heterogeneity is the key to long living shear zones it is not clear whether heterogeneity is embedded primarily at brittle level thus repeatedly triggering ductile shear zones in corresponding locations at depth or whether it is caused by structural modification in the ductile field itself. We would argue here on the basis of the low strength of the brittle zone (Fig. 7) and the proclivity of the strong semi-brittle layer to nucleate thermal–mechanical shear zones that brittle fault zones play a minor role in engraving plate boundaries over long time scales. They play, however, an important role in the earthquake cycle.

7. Towards earthquake modelling

Comprehensive numerical approaches to earthquakes require a simultaneous solving of multiphysics feedback processes at mm scale and a consideration of the dynamic changes at large geodynamic scale. The important issue of coupling tectonic and seismic length scales has already been pointed out 10 years ago (Anders and Sleep, 1992). In spite of the rapid evolution of computational power, we have not reached the required temporal and spatial resolution to do this. However, a strong economical push towards solving, amongst other geodynamic phenomena, the earthquake problem has led to the development of large-scale Earth computational projects such as ACcESS (http://www.quakes.uq.edu.au/) investigating the application of classical and new numerical techniques. We will briefly summarize the current state of this rapidly developing field with the most elementary approaches.

7.1. Brittle models

One-dimensional brittle earthquakes models have been formulated analogous to the ductile earthquake model discussed earlier. A constant velocity is applied to a simple spring–slider block system where the sliding friction of the block replaces the ductile flow in the shear zone. The friction law can be derived from laboratory data giving a friction coefficient that is dependent on the velocity and on the contact state, e.g. gouge layer, between the sliding surfaces (Dieterich, 1979b, 1992). Later work included also the effect of shear heating (Blanpied et al., 1998; Chester and Higgs, 1992) in the constitutional rate and state variable friction model (Kameyama and Kaneda, 2002). Kato (2001) showed that the shear heated fault goes unstable at about 20% smaller pre-seismic sliding compared to the fault without shear heating. The style of instability is the same as shown in Fig. 10, i.e. from observational data consistent with seismological studies of earthquake rupture characteristics, there is no clear difference between a ductile and a brittle 1D earthquake.

Two-dimensional models of brittle earthquakes have also used a simplified Coulomb failure analysis in which the dynamics of the rate and state variable friction is neglected and the effective friction is governed by the Coulomb failure envelope. It is assumed that the predefined fault remains locked during loading until it reaches its failure criterion and then it fails instantaneously. In between the faults, the material is assumed to interact elastically. Both Coulomb and rate and state variable friction models (Tullis, 1999) can be
applied to data from modern earthquake catalogues, however, there is no clear evidence of superiority of one concept over the other (Gomberg et al., 2000). While the potential of identifying faults from geodetic observations and earthquake catalogues makes this style of analysis appealing, the restriction to failure on predefined faults without their capacity of developing smaller scale faults or dynamic evolution of friction is a severe limitation. It has been shown that a complex fault network displays dynamical modes not observed in simple fault systems (Rundle et al., 2001). The behavior of the entire earthquake fault network system appears to self-organize in space and time into particular modes that are also controlled by the interaction of changes in physics on the scale of single faults and smaller (Ben-Zion and Sammis, 2003).

The fault network thus must be modeled as a whole and the potential of fault zone propagation, degradation and healing must be built into the constitutive law with considering a full coupling to the energetics. Methods that allow just this have been presented for both the brittle and the ductile field. They have been found to be successful for describing the longevity and memory of fault zones. An excellent discussion of the brittle damage mechanics approach to model single and network fault system has been given in Lyakhovsky et al. (2001).

Another relatively new numerical approach has the same potential. It is an up-scaled version of molecular dynamics calculations. Rather than formulating the mathematical problem in terms of a continuum it is reduced to calculating the interaction between discrete particles which when put together mimic the physics observed at larger scale. Currently, friction, fracture, granular dynamics and thermal–mechanical and thermal-porous feedback have been implemented (Abe et al., 2000). An example of granular dynamics modeled with the particle approach is shown in Fig. 18. Since the approach places itself at the lower scale of the upscaling scheme in the Table 1, it has the advantage that micro-scale physics that are potentially overlooked in the larger scale approaches are not neglected. The obvious disadvantage is that geodynamic scales cannot yet be reached owing to numerical constraint. The approach is not restricted to the brittle field but ductile shear zones can also be modelled, using a discrete particle approach (Li and Liu, 2000; Mora and Place, 1998).

Fig. 18. Discrete particle-dynamics calculation of a granular shear zone (Mora and Place, 1998). A normal stress of 150 MPa is maintained on the upper and lower edges of elastic blocks above and below the region. The block is sheared as indicated by the arrows and the deviatoric stress is monitored showing filamentary paths with high stress. The model can explain the so-called “heat flow paradox” in the brittle crust of the San-Andreas Fault and is a good example for self-organized brittle network sketched in Fig. 19.

7.2. Brittle versus ductile earthquakes

We have not discussed shear zone formation due to phase transformations because of their restriction to limited $p-T$ conditions. Phase transformations may not be capable of supplying a universal ductile earthquake mechanism but they may play a role in preparing conditions for deep ductile earthquakes (Karato et al., 2001). However, we have pointed out that outside the $p-T$ conditions necessary for olivine-spinel transformations there is already one important mechanism for ductile earthquakes relying on thermal–mechanical feedback. We have shown visco-elastic (Ogawa, 1987), elasto-plastic (Hobbs and Ord, 1988; Roberts and Turcotte, 2000) and elasto-visco-plastic (Regenauer-Lieb and Yuen, 2000a) ductile thermal–mechanical earthquake mechanism. The question arises as to whether we can discriminate between brittle and ductile earthquakes from observational data (Wiens and Sneider, 2001).

One important observational constraint is the direct or indirect observation of heat released during an earthquake or the cumulative heat released during
prolonged seismic activity. The observation of negligible heat flow anomaly over the San Andreas fault zone at the Cajun Pass has been publicized as the “heat flow paradox” (Scholz, 2000a,b). The brittle granular calculation of Mora and Place (Mora and Place, 1998) has indeed proven that there is nothing paradoxical about low heat flow in a granular shear zone. We have pointed out that there is some thermal mechanical feedback to be expected also in the brittle field—in fact, it is possible to come up with a frictional theory that relies on temperature (Kameyama and Kaneda, 2002)—a good indicator for brittle shear zones is their lower strength and consequently also their lower heat release than their ductile counterpart. We have pointed out that brittle fault zones are also subject to degradation or healing, hence, the opposite case of relatively high heat release is no unequivocal evidence of ductile earthquakes.

Another indirect evidence for thermal–mechanical feedback is the observation of collocated deep earthquakes repeating in the same area within days (Wiens and Snider, 2001). Thermal diffusion on a thin, meter-scale, thermal–mechanical shear zone provides a viable mechanism for repeating earthquakes. When including thermal-elasticity into our numerical calculation for subduction initiation (Regenauer-Lieb et al., 2001) we obtained thermal–mechanical instabilities that comprise one element size showing that thermal–mechanical ductile earthquakes are indeed expected to have very narrow fault planes. We would like to point out that only a modest amount of about 20 K (Roberts and Turcotte, 2000) shear heating is necessary to turn aseismic creep into a seismic instability. Therefore, we conclude that ductile earthquakes belong as a natural element to some mylonitic shear zones. Whether the ductile instability turns seismic or whether there is just a phase of accelerated creep (Ben-Zion and Lyakhovski, 2002) depends on the temperature and material parameter in the shear zone.

8. Summary

We have been discussing the basic numerical shear zone concepts, their potentialities and their limits. Thermal–mechanical shear zone formation has been shown to rely on momentum– and thermal–mechanical feedback processes. While the importance of thermal–mechanical feedback in the brittle field is weak, leading to the acceleration of the onset of seismic instabilities, seismic instabilities or formation of shear zones in the ductile field rely intrinsically on thermal mechanical feedback fed by the exponential dependence of creep strength on temperature. When a shear zone has been fully developed, the relative role of feedback processes changes. Deformational work, dissipated prior to the formation of the shear zone in a continuum around the shear zone, is now released within the shear zone. This leads to important modifications of the energetics of the faulted system. We have pointed out the implications of diverse time and length scales.

On the large plate tectonic time-space scale, the following characteristics have been derived. Mylonitic shear zones take over the mechanical control of the whole lithosphere. During the evolution of the deformation viscous modeling shows that mylonitic shear zones become continually weaker, owing to the increasing temperature inside the shear zone. This temperature increase would go to a quasi-steady state value that depends on the thermal properties of the sheared lithology and its activation energy (Eq. (35)) and does not exceed 100–300 K reached after 10 Ma shearing (Fleitout and Froidevaux, 1980). The width of the predicted thermal shear zone (about 20 km) is much larger than observed large-scale mylonite shear zones of the scale of a few kilometers (Hobbs et al., 1986). To resolve this discrepancy, other modifications of the energetics have been considered.

Grain size sensitive creep can only be efficient under a narrow parameter range of shear zone cooling (Braun et al., 1999). This is not possible with positive shear heating but is workable during an intermediate

phase of cooling after a ductile earthquake (Fig. 10) or if uplift or fluids cool down the shear zone. We can assess the significance of grain size sensitive creep for shear zone formation on the basis of a simple function of the cooling rate for localization by grain size sensitive creep given by Braun et al. (1999) in SI units:

Cooling rate for grain size sensitive shear zones

$$\log_{10} \left( \frac{\dot{T}}{\lambda} \right) = \log_{10} \dot{\varepsilon} + 1.7$$  \hspace{1cm} (40)

where $\dot{\varepsilon}$ is the constant defined in Eq. (17) and has a value that lies between 10 and 20. Note that this approximation has been derived by neglecting thermomechanical coupling (Kameyama et al., 1997) and the non-linear (power-law) aspect of the flow law (Solomatov, 2001). However, when applying the criterion to observed shear zones, the restrictive nature of thermomechanical boundary condition for localization by grain size sensitive creep becomes apparent.

Individual fault segments inside mylonitic shear zones have a width that lies well below 1 m (Drury et al., 1991). Hence, if a width of 1 m were controlled by thermomechanical conditions, we would imply a strain rate for grain size sensitive creep larger than $10^{-6}$ s$^{-1}$ (Fig. 19). From Eq. (40), we obtain a cooling rate that must be of the order of $10^{-3}$ K s$^{-1}$. Such conditions are only possible after a ductile

---

**Fig. 19.** Synopsis of shear zone observations in the field and inferred material properties and length scales for modelling to the left and right sides of a network of brittle fault zone (top) and ductile mylonitic shear zones (bottom). In the continental crust, the onset of quartz intracrystalline plasticity at about 270 °C (van Daalen et al., 1999) marks the transition from fault to (semi-brittle) mylonitic shear zones.
earthquake (Fig. 10). Ductile earthquakes, or phases of accelerated aseismic creep $>10^{-12}$ s$^{-1}$ would indeed predict sub-kilometer scale shear zones as observed in the field. The magnitude of the heating pulse during a ductile event (shown in Fig. 10 lies well outside the conditions necessary to leave tell-tale melts behind, so-called pseudo-tachylites). Grain size reduction alternating with ductile earthquakes are a viable explanation for large-scale networks of mylonitic shear zones (Jin et al., 1998; Montesi and Hirth, 2003). This would imply a jerky flow at the scale of the shear zone and the time scale of several thousand years. Jerky flow is not uncommon in nature. For instance, also found in metals at much smaller time and space scales (Lebyodkin et al., 2001).

It still remains an open question as to whether the small-scale thermal—mechanical conditions inside the individual shear strands can control the large-scale behavior of the entire shear zone. In order to resolve this question, we need more powerful numerical techniques that are able to resolve locally in centimeter scale and at the same time consider large-scale geodynamic boundary condition at 1000 km scale. Adaptive wavelet-based techniques (Vasilyev et al., 2001) have the potential to do this and they may in the future displace finite element approaches. Since the thermal wave disappears over geological time scale and is larger than observed large-scale mylonitic shear zone (e.g. the Redbank shear zone in Australia (Hobbs et al., 1986)) we have argued that longevity and memory of shear zones must rely on additional non-thermal storage of energy dissipated inside the shear zone.

For a shear zone to become geologically permanent, we should consider energy storage in terms of new surface area as it is given by the nucleation of volatile filled voids (Bercovici et al., 2001a; Regenauer-Lieb, 1999). Observations on volatiles released from mantle shear zones shows that the maximum width of the degassing zone is about 10 km (Regenauer-Lieb, 1999). This volatile rich zone thus constrains the largest size possible for a mylonitic shear zone. Experiments with rock analogues (Bauer et al., 2000; Bons et al., 1993) and real rocks (Dieterich, 1979a; Mandl et al., 1977; Post, 1977) as well as engineering applications (Ananthakrishna et al., 2001; Fressengeas and Molinari, 1987). We have outlined the recent advances in numerical modeling of shear zones and have emphasized the multiscale physics of feedback mechanisms that are important. In the synopsis we have pointed out how geological observations of shear zone length scales can be helpful in interpreting the basics of shear zone properties with the aid of the simple parametric laws that are obtained from numerical modeling. The next step forward would be to go beyond the specializations inherently gained in the various fields and compile a truly multidisciplinary dataset useful for lithosphere dynamics (Handy et al., 2001).
Acknowledgements

We thank Bruce Hobbs, David Bercovici, Yehuda Ben-Zion, Peter Mora, Hans Muhlhaus, Slava Solomonov, Shun-Ichiro Karato and Charley Kameyama for stimulating discussions. This project has been supported by both the geophysics program of the N.S.F. and the Swiss Nationalfond 21-61912.00. This is publication 1231 of the ETH Zurich.

Appendix A. From triaxial experiment to triaxial flow law

A.1. Associated flow law

Because of the large experimental uncertainties, the appropriate transformation of laboratory to tensorial creep laws has only been discussed parenthetically (Nye, 1953; Ranalli, 1995). If we assume that the material is isotropic throughout flow and incompressible, a generalized flow law can be expressed. We first describe the classical purely plastic flow law:

\[
\dot{e}^{pl}_{ij} = \lambda \sigma_{ij}
\]  

(A1)

where the superscript pl refers to plastic strain rate and \( \lambda \) is a function of position and strain history. In plasticity theory, it is not a material property but a scalar multiplier with dimension \( \text{s}^{-1} \text{ Pa}^{-1} \), which is zero when the stress state is below the yield stress (e.g. in the inside of the cylinder shown in Fig. 3) and some positive value corresponding to the strain-dependent hardening law allowing the cylinder to grow or shrink as a function of strain hardening or weakening, respectively. This is known as the Levy–Mises flow law. It states that the stress and strain rates are everywhere co-axial meaning that the principal axes of the stress tensor and the strain rate are coincident. The flow law furthermore implies that the components of strain rate are proportional to components of the deviatoric stress only and there is no pressure sensitivity. The classical theory of plasticity does not consider time as a degree of freedom and therefore the Levy-Mises flow law is originally formulated with respect to the strain increment tensor instead of the strain rate tensor as illustrated in Fig. 3 in order to emphasize the time invariance. In this figure, the principal of “normality” is also illustrated implying that the principal stress, strain increment and strain rate axes, are normal to the yield cylinder. Whenever we refer to this style of flow law it is called “associated plasticity” synonymous with “coaxial flow” or the flow is also said to be “normal” to the yield envelope.

In classical linear fluid mechanics, the same coaxial flow law is used but time plays a role, although explicit time-dependent solutions can often be avoided due to extremely slow, so called creeping flow where the energy equation sometimes does not need to be solved (see comments applied to creeping flow in Appendix C). There is no yield criterion and \( \dot{\lambda} \) becomes a true material property (the inverse of viscosity), being constant for the simple Newtonian fluid. Here, we are dealing with more complex flow laws, which have a non-linear stress versus strain rate relationship. In order to extend Eq. (A1) into an associated flow law that has a non-linear stress–strain relation, it is convenient to introduce scalar measures of deviatoric stress and strain rate. Following Nye (1953), we have defined an effective stress and an effective strain rate, in Eqs. (3) and (5) accordingly. Nye’s formulation is motivated by a pure shear plane strain experiment in which the intermediate principal strain rate is zero and continuity requires that the maximum and minimum principal strain rates are equal but have an opposite sign.

\[
\sigma' = \sqrt{\frac{1}{2} \sigma_{ij} \sigma_{ij}} \\
= \sqrt{\frac{1}{6} \left[ (\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 \right]} 
\]

(A2)

\[
\dot{\varepsilon} = \sqrt{\frac{1}{2} \dot{\varepsilon}_{ij} \dot{\varepsilon}_{ij}} \\
= \sqrt{\frac{1}{6} \left[ (\dot{\varepsilon}_1 - \dot{\varepsilon}_2)^2 + (\dot{\varepsilon}_2 - \dot{\varepsilon}_3)^2 + (\dot{\varepsilon}_3 - \dot{\varepsilon}_1)^2 \right]} 
\]

(A3)

the indices 1, 2, 3 refer to the principal stresses and strain rates, respectively, and \( \sigma_{ij} \) is the deviatoric stress tensor. Note that the effective stress and strain rates are always positive. The above definition ensures
that the effective strain rate is equal in magnitude to
the maximum or the least principal strain rate measure
in the pure shear experiment.

It is straightforward to reformulate Eq. (A1) now
for a more generalized associated flow law where the
scalar factor is a function of the effective stress.

\[ \dot{e}_{ij}^{\text{visc}} = \frac{f(\sigma')}{\sigma'} \sigma_{ij} \]  

(A4)

where the superscript visc now refers to viscous strain
rates. For example, the power-law formulated in Eq.
(5) implies a tensorial viscous co-axial flow law

\[ \dot{e}_{ij}^p = a^n \sigma^{(n-1)} \sigma_{ij} \]  

(A5)

This viscous flow law turns into a visco-plastic flow
law if we define a yield threshold like in the Bingham
formulation (Eq. (8)). If we allow in addition elastic
deformation before reaching the yield threshold (Eq.
(A1)), we obtain an elasto-visco-plastic flow law.

Nye’s (1953) definition of effective stress and
strain rate is adopted in Ranalli’s textbook (Ranalli,
1995). We suggest to use a slightly different notation,
popular in the engineering community (Chakrabarty,
2000), which is motivated by triaxial conditions
depicted in Fig. A1 instead of the pure shear con-
ditions in the classical definition.

\[ r = \sqrt[3]{2} \sigma' \]  

(A8)

\[ e = \sqrt{\frac{3}{2} \dot{e}_y \dot{e}_y} = \sqrt{\frac{1}{2} \left[ (\dot{e}_1 - \dot{e}_2)^2 + (\dot{e}_2 - \dot{e}_3)^2 + (\dot{e}_3 - \dot{e}_1)^2 \right]} \]  

(A9)

\[ \dot{e}_1 + \dot{e}_2 + \dot{e}_3 = 0 \]  

(A6)

Because of rotational symmetry of the experiment
around the maximum compression axis, the interme-
diate and least principal radial strain rates \( \dot{e}_2 \) and \( \dot{e}_3 \) are
of equal magnitude and it follows that their magnitude
is half the axial strain rate.

\[ \dot{e}_3 = \dot{e}_2 = -\frac{1}{2} \dot{e}_1 \]  

(A7)

For ease of implementing laboratory data, we use,
however, a slightly different formulation based on the
von Mises equivalent deviatoric stress of the experi-
ment. These are obtained from Nye’s original for-
mulation of Eqs. (A2) and (A3) by a scalar multipli-
cation with the square root of three.

A.2. Triaxial experiment

Laboratory experiments usually report the creep
law in terms of differential stress versus uniaxial strain
rate in the piston direction of a triaxial experiment.
The experiment is shown in Fig. A1.
Using the convention that compressive strain rates and stresses are positive, we obtain the axial stress in Fig. A1 as the maximum principal (compressive) stress $\sigma_1$ and the radial stress of the confining medium constitutes $\sigma_2 = \sigma_3$. It follows that the effective stress for the numerical implementation can be calculated from the laboratory deviatoric stress $\sigma_D = \sigma_1 - \sigma_3$ as:

$$\sigma' = |\sigma_D|$$

(A10)

By analogy, the effective strain rate can be computed from the axial strain rate reported in the experimental flow laws.

$$\dot{\varepsilon} = \frac{3}{2} \dot{\varepsilon}_1$$

(A11)

The only factor is thus the constant 2/3 for converting laboratory flow laws into effective flow laws for numerical calculations. This factor is valid for any flow law. Note that Nye’s original formulation of effective stress and strain rate based on pure shear is awkward for rescaling triaxial experimental results into effective flow laws.

This has also been noted by Ranalli. For power-law, for instance, the strain rates have to be multiplied by a factor of $2/(3(n+1)/2)$ to transform triaxial deviatoric stress–uniaxial strain rate equations into Nye’s effective quantities (Ranalli, 1995). If this rescaling is neglected, an increasingly large error is implied for increasing $n$, e.g. one order of magnitude for $n = 4.5$. For the Peierls stress, yet another scaling factor is required, which is obsolete, if the above triaxial definition of effective stress and strain rate is chosen.

Appendix B. Non-associated flow laws and localization

For the mathematical treatment of mylonitic shear zones, we have been dealing with associated flow laws described in Appendix A for which the energy theory of localization is required. We have, however, also discussed a dilatant plastic material, which turns into a strongly non-associated material when the brittle void nucleation criterion is used (Eq. (27)). For details, see Needleman and Tvergaard (1992); this paper also provides an excellent review of the constitutive theory of localization, which is a suitable criterion for describing localization phenomena in the brittle field.

In the following, we are giving a brief introduction into non-associated plasticity, using the example of the Mohr–Coulomb criterion. A very detailed review of non-associated plasticity is found elsewhere (Vermeer, 1984). Subsequently, we briefly discuss the classical bifurcation analysis and the development of a hardening law that can lead to bifurcations. Note that the constitutive theory of localization has not been developed to comprise localization of strain rate and thermally sensitive solids (Rice, 1977). The following discussion therefore interprets the flow law in terms of a time-independent plasticity criterion only.

B.1. Non-associated plasticity and corners in the yield envelope

If we neglect time-dependent quantities, Eq. (31) simplifies into and elasto-plastic body

$$\dot{\varepsilon}_{ij} = \dot{\varepsilon}_{ij}^E + \dot{\varepsilon}_{ij}^p$$

(A12)

where we only add elastic and plastic strain rates. The transition from a purely elastic to an elasto-plastic state is given by the yield envelope. Eq. (23) gives an example of an elasto-plastic pressure and deviatoric stress-dependent yield function $\Phi$ which collapses into the von Mises envelope for $q_{1,2,3} = 0$. The von Mises yield envelope is illustrated in Fig. 3, which is the basis for definition of the scalar multiplier $\lambda$ in the associated Levy–Mises flow law, i.e. $\lambda = 0$ inside the von Mises cylinder and on the cylinder $\lambda > 0$.

$$\Phi = \left(\frac{\sigma'}{\sigma_y}\right)^2 - 1 = 0$$

(A13)

For a von Mises solid, the direction of flow is normal to the yield surface, so the flow potential coincides with the yield envelope and the flow law can be expressed as a function of the effective stress only. Extending the flow law into a generalized plastic flow law where flow and yield potential may not coincide, we rewrite Eq. (A1) into

$$\dot{\varepsilon}_{ij}^p = \lambda \frac{\partial G}{\partial \sigma_{ij}}$$

(A14)

where $G$ is the flow potential, giving the direction of flow after yielding. If $G = \Phi$, the flow is associated but
for inequality, we speak of a non-associated flow, i.e. the flow is not normal to the yield envelope. An example for non-associated flow is the Mohr–Coulomb plastic body, which is here expressed in terms of principal stresses (Vermeer, 1984):

$$
\Phi = \frac{1}{2}(\sigma_1 - \sigma_3) + \frac{1}{2}(\sigma_1 + \sigma_3)\sin\mu - c\cos\mu \quad (A15)
$$

where $\mu$ is the friction angle from the Mohr-Coulomb failure envelope and $c$ the cohesion. This yield envelope has corners when plotted in the three-dimensional stress space. Another feature of the Mohr–Coulomb failure envelope is that it does not depend on the intermediate principal stress as shown in Eq. (A15). Similar to the yield envelope the flow potential does not depend on the intermediate principal stress but depends on the dilatancy angle $\beta$ instead of the friction angle $\mu$.

$$
G = \frac{1}{2}(\sigma_1 - \sigma_3) + \frac{1}{2}(\sigma_1 + \sigma_3)\sin\beta + \text{const} \quad (A16)
$$

if the dilatancy angle $\beta$ is equal to the friction angle $\mu$, the Mohr–Coulomb law turns into an associated flow law. This condition is, however, too restrictive ( Rudnicki and Rice, 1975).

Interpreting the Mohr–Coulomb flow law in the framework of the triaxial experiment ($\sigma_2 = \sigma_3$) shown in Fig. A1, it becomes immediately apparent from substituting $\sigma_3$ by $\sigma_2$ in Eq. (A16) that we have to deal with two potential functions and two scalar factors owing to a corner on the yield envelope.

$$
\dot{\varepsilon}_{ij}^{pl} = \dot{\lambda}_1 \frac{\partial G_1}{\partial \sigma_{ij}} + \dot{\lambda}_2 \frac{\partial G_2}{\partial \sigma_{ij}} \quad (A17)
$$

It is obvious that rheologies with such singular transitions in flow laws are prone to localization on preferred planes.

B.2. Localization bifurcations

For a quantitative investigation of these instabilities, bifurcation analyses have been done (Needleman and Tvergaard, 1992; Rice, 1977) and critical hardening has been predicted as a basis for a constitutive theory of localization. In such analyses, conditions for band like instabilities within homogeneous, homogeneously deforming rate-independent solids are derived and a correspondence for the occurrence of stationary body waves has been found (Rayleigh and Stoneley waves). Characteristic directions of localization are found to be preceded and guided by these elastic wave phenomena. Bifurcations are associated with a loss of ellipticity in rate-independent solids. However, for rate-dependent solids, the constitutive theory finds that localization bifurcations are effectively suppressed, i.e. the governing equations remain elliptic. The energy theory of localization provides further insight for the case of ductile shear zones.

**Appendix C. Energy theory of localization**

As an additional element, the energy theory of localization takes the modification of the local energy during deformation into account. This is a very important step up in physics and requires marriage of classical mechanics and non-equilibrium thermodynamics. Note that both the continuity and the momentum equations are independent of time. Time dependence only arises through the energy equation. We have already pointed out that, in geodynamics and engineering applications, we can often use a quasistatic (plasticity theory) or the so called creeping flow regime (fluid mechanics), where it is possible to sometimes ignore the effect of time. However, whenever throughout deformation, there appear time derivatives in the local energy quantities, the time invariance must be abandoned. This applies to both solid- and fluid-mechanics, although it appears to be more naturally accepted in the mantle convection community, for which reason most of the early energy concepts to localization, reviewed here, have been dealing with fluid rheologies.

It is obvious that solids display some important modifications in the local energetics after shear zone formation. It follows that the energy theory is required for defining the shear zone width after localization. The lack of a physical scaling length governing shear zone width is one of the shortcomings of the classical constitutive theory of localization. We will show that a critical local energy density acts as a trigger to localization, thus overcoming the elliptical solution space for ductile rate-dependent solids, which is the second most spectacular failure of the classical constitutive theory of localization.
C.1. Thermodynamic criteria for instability

For this, we look at the internal energy of a homogeneous volume element of a sample, which in classical equilibrium thermodynamics is characterized by \( n + 1 \) state variables, where temperature is the variable for \( n = 0 \). For the time being, we assume that there is no flux of energy from external sources through a radiation term. It has been argued that this is not a necessary restriction since the radiation term does not determine the thermodynamic process but, rather, the thermodynamic process determines it (Lavenda, 1978). The book by Lavenda gives a critical review of the theory of non-equilibrium thermodynamics from the time of birth of the chaos theory. We recommended this book as a further reading on the basic concepts. In the following, we only give a brief introduction to the theory.

During mechanical deformation thermodynamic state variables, governing mechanical properties comprise first of all the elastic strain and the absolute temperature. During deformation, additional microstructural variables come into play, which can characterize dislocation density, phase changes, damage (new surface energy), etc. These are often expressed in terms of tensorial functions of strain rate energy densities (product of stress and strain rates). We have discussed that before flow bifurcation the system deforms homogenously so that non-equilibrium localization phenomena can be seen as a dynamic sequence of evolving thermodynamic equilibrium states. Taking one particular equilibrium state at time \( t \), we write the specific Helmholtz free energy \( \psi \) of this volume element as a function of its \( n + 1 \) state variables

\[
\psi(T, e^E, x_i), \quad 1 \leq j \leq n
\]

where the elastic strain \( e^E \) and the absolute temperature \( T \) are the first two state variables \( x_{0,1} \). The second law of thermodynamics leads to the inequality of Clausius–Duhem

\[
-\frac{D\psi}{Dt} - \frac{q}{T} \nabla T = \sigma_{ij} \dot{e}_{ij} - \rho \frac{\partial \psi}{\partial x_j} \frac{Dx_j}{Dt} - \frac{q}{T} \nabla T \geq 0
\]

(A19)

where \( q \) is the heat flux vector out of the reference volume and the term including the material time derivative \( Dx_i/Dt \) gives the stored energy terms, which appears for instance in new surface energy during microcracking (Chrysochoos and Peyroux, 1997). Now, the double product of the Cauchy stress tensor and the strain rate tensor gives the mechanical power, which also contains the non-dissipative reversible elastic deformational work rate (so-called isentropic power). We subtract this work out of the entropy change in Eq. (A19) by the first material derivative of the specific Helmholtz free energy for \( n = 1 \). We already would like to point out here that an additional feedback term (comprising the second derivative) appears later in the derivation of the energy equation. By analogy, all other stored energy terms for higher \( j \) can be subtracted likewise. In order to assess the dissipation out of equilibrium, we define the intrinsic dissipation function \( R_i \) and perturb the equilibrium system by small velocity perturbations \( \xi_i \).

\[
\frac{1}{2} R_i \xi_i \xi_j = -\frac{D\psi}{Dt} - \sigma_{ij} \dot{e}_{ij} - \rho \frac{\partial \psi}{\partial x_j} \frac{Dx_j}{Dt}
\]

(A20)

Eq. (A19) embodies the core of the energy theory of localization, mathematically expressing the dissipation as the sum of force-flux products where within each product there appears a state variable. While in classical plasticity, the system is considered mathematically closed when the conservation laws of mass and momentum are satisfied, in the energy approach the additional consideration of the energy fluxes in the specific entropy production associated with \( R_i \) gives a closed system. This opens the way to non-elliptical solutions essential for the phenomenon of localization (Aifantis, 1987) as we show below.

A necessary but not sufficient condition for stability is that the system dissipates positively. This implies that in a generalized Gibbs space, spanned by all independent velocities, the dissipation function must be given by an ellipsoid centered on the origin. Following states can be distinguished in a generalized velocity space (Lavenda, 1978):

- **Ellipsoid**, for all \( i \quad R_i > 0 \)
- **Parabolic**, at least one \( R_i \geq 0 \) (A21)
- **Hyperboloid**, at least one \( R_i < 0 \)

The ellipsoid space is a necessary but not a sufficient condition for a homogeneous solution. Within the ellipsoid solution space geometrically controlled shear...
zones are possible, for instance. We will not deal with such shear zones that are predefined by geometry. The parabolic regime, on the contrary, is a sufficient condition for material instability and it is said to be in a state of meta-stability. In thermodynamics, this is also called a “marginally stable state”. It follows that a single internal state variable linked to a source of internal power can cause flow localization (Fig. A2).

![Fig. A2. Localization criteria and their expression in a pure shear experiment. $R_i$ is the intrinsic dissipation function defined in Eq. (A20). Without feedback ductile, deformation takes place in the elliptic regime where homogeneous deformation persists. Shear localization on preferred slip planes arises when one of the $R_i$’s, related to a single thermodynamic state variable, is zero. The energy theory offers a closed mathematical system to calculate these bifurcations. Intrinsic length scales, defining critical energy levels, ensure that solutions stay in the parabolic regime, i.e. shear bands have a finite thickness. There is no mesh sensitivity if the numerical resolution matches intrinsic length scales. In the hyperbolic regime, shear zones become slip lines, i.e. they are mathematical idealization with a vanishing thickness. The method of characteristics (Dewhurst and Collins, 1973) is a suitable mesh-less solution tool for these idealized rheologies.](image-url)
While this terminology defines flow localization within the framework of the energy theory, we also would like to be able to predict post-bifurcational evolution of the shear zone. The conservation law of power introduced in the following paragraph allows the incorporation of feedback terms, while the variational principle of least dissipative power gives post-bifurcational time-dependent evolution, which is self-consistent in terms of classical mechanics and compatible with non-equilibrium variational thermodynamics. This enables us to formulate criteria for numerically tractable solutions in the post-bifurcational state using the variational principles of finite element analyses and an adaptive time stepping scheme controlled by a critical thermodynamic meta-stable state (Regenauer-Lieb and Yuen, 2003).

C.2. Energy equation

Consider the same volume element in thermodynamic equilibrium. We have been describing its internal power in motion using the Lagrangian, also called the substantial or material time derivative implying that in our mathematical description we are moving with the deforming volume element. In fluid mechanics, this volume element is sometimes called a fluid parcel. Integrating with respect to time, we thus obtain the specific internal energy \( e_{\text{int}} \) of the reference volume/parcel in motion and considering its kinetic energy \( e_{\text{kin}} \) by inertia we obtain with the classical mechanical energy balance.

\[
\int_V \rho e_{\text{tot}} \, dV = \int_V \rho e_{\text{int}} \, dV + \frac{1}{2} \int_V \rho \mathbf{v} \cdot \mathbf{v} \, dV \quad (A22)
\]

where \( \mathbf{v} \) is the velocity vector and \( V \) the reference volume. We are interested in how this energy changes with time so in the following we always consider the substantial/material derivative and the law of energy conservation turns into a conservation law of power.

\[
\frac{D e_{\text{tot}}}{D t} = \frac{D e_{\text{int}}}{D t} \quad (A23)
\]

The thermodynamic energy balance for the specific energy is given in terms of entropy \( s \) by

\[
e_{\text{int}} = \psi(T, \epsilon^{\text{el}}, \mathbf{z}_j) + sT \quad (A24)
\]

Additionally

\[
s = -\frac{\partial \psi}{\partial T} \quad (A25)
\]

and

\[
\frac{D s}{D t} = -\frac{\partial^2 \psi}{\partial T^2} \frac{D T}{D t} - \frac{\partial^2 \psi}{\partial T \partial z_j} \frac{D z_j}{D t} \quad (A26)
\]

where the specific heat \( c_a \) is defined as

\[
c_a = -T \frac{\partial^2 \psi}{\partial T^2} \quad (A27)
\]

In the development of the specific Helmholtz free energy of the reference volume, we have assumed that the flux of power \( r \) by radiation is zero. We now relax this condition and write the basic balance of power, which in continuum thermodynamics is given (Green and Naghdi, 1965)

\[
\int_V \rho \frac{D e_{\text{int}}}{D t} \, dV = \int_A q \, dA + \int_A r \, dA \quad (A28)
\]

where \( A \) now stands for the surface area of the reference volume \( V \) and outwards directed flux is positive. Substituting Eqs. (A24–27) into Eq. (A28), we obtain

\[
\int_V \rho \frac{D e_{\text{int}}}{D t} \, dV = \int_A q \, dA + \int_A r \, dA \quad (A29)
\]
Using the intrinsic dissipation defined in Eq. (A20) and rearranging terms the balance of energy now reads:

\[
\int_V \rho c_2 \frac{DT}{Dt} \, dV = \int_V \sigma_{ij} \dot{\varepsilon}_{ij} - \rho \frac{\partial \psi}{\partial x_j} \frac{Dx_j}{Dt} \\
+ \frac{\partial \partial \psi}{\partial T \partial x_j} \frac{Dx_j}{Dt} \, dV - \int_A q \, dA \\
- \int_A r \, dA \tag{A30}
\]

When comparing Eq. (A30) with Eq. (A19), we note a new, third term on the right side of Eq. (A30). This gives the additional coupling term in the energy equation, while the last term including \( r \) is the external source term through e.g. radiation, chemical reactions and Joule heating, etc.

The energy Eq. (A30) is completely based on thermodynamic state variables; we will now go on and simplify. This is done by considering, what we appreciate to be, the most important effects. Note that there is no current consensus on the role of elasticity between fluid and solid-mechanical communities and the feedback owing to the creation of new surface energy (void creation) may be considered more important than the effect of elasticity. We have argued in this review that both terms are important in the ductile damage through void creation summarized here is severely simplified. A more complete thermodynamically inspired theory of localization due to void creation is therefore introduced as the elastic strain and this coupling term consequently describes the thermal-elastic effect. Thermal-elasticity takes into account that the material dilates on heating and shrinks on cooling. Another important coupling could be latent heat release upon phase transitions. In the following, we simplify the energy equation by only writing down the elastic coupling term, which is also known as isentropic power by adiabatic volume changes being

\[
\rho T \frac{\partial^2 \psi}{\partial x_j \partial x_j} \frac{Dx_j}{Dt} = \lambda \, T_\text{equ} \, \frac{Dp}{Dt} \tag{A31}
\]

where \( \lambda \) is the linear coefficient of thermal expansion and \( T_\text{equ} \) is the equilibrium temperature change of adiabatic expansion/compression.

We use the additive Maxwell body decomposition (e.g. Eq. (A12)) and note that we can separate out visco-plastic from elastic power by:

\[
\sigma_{ij} \dot{\varepsilon}_{ij} = \sigma_{ij} \dot{\varepsilon}_{ij}^E + \sigma_{ij} \dot{\varepsilon}_{ij}^\text{visc} \tag{A32}
\]

We have already discussed the influence of the elastic power, now we will discuss the second term, the double product of visco-plastic strain rates and the stress tensor, giving the dissipative power:

\[
\sigma_{ij} \dot{\varepsilon}_{ij}^\text{visc} = \chi \sigma_{ij} \dot{\varepsilon}_{ij}^\text{visc} + \gamma (p - p_\nu) \frac{1}{\rho^2} \frac{\partial \rho}{\partial t} \tag{A33}
\]

where we separate out deviatoric from isotropic dissipation processes. The prefactors give an additional simplification by dropping the stored energy terms (e.g. surface energy due to dilatancy which would go into the second term on the right in Eq. (A30)) and lumping them into a scalar factor \( 0 < \gamma, \gamma < 1 \) thereby diminishing the shear heating term or the dissipation through the volume change, respectively. The pressure stress \( p_\nu \) is due to bulk viscosity causing the total dissipative volume change. This term is often considered in extended Boussinesq approximations of mantle convection (Yuen, 2000) but for the purpose of
faulting in the ductile part of the lithosphere it can be conveniently neglected. The energy equation is now simplified to

$$\frac{\partial T}{\partial t} = \frac{c_p}{\rho} \frac{\partial T}{\partial t} = \chi \sigma_{ij} \dot{e}_{ij} + \lambda \frac{\partial T}{\partial t} - \rho c_p \kappa \nabla^2 T$$  

\hspace{1cm} (A34)

where the conduction is spelt out in terms of diffusivity $\kappa$ (note that, for large strain conduction, it is also strain-dependent (Povirk et al., 1994) and not necessarily isotropic).

C.4. Simplified energy theory of localization in Earth and engineering sciences

While the constitutive theory appears to have reached a stage of maturity, the energy theory definitely has not. In particular, the vexing separation of the different scientific communities and the differences in notation of geo- and engineering style has prohibited its level of acceptance. Perhaps, the most radical drawback in Earth Sciences is that the brittle solid observed at surface localizes readily, so that little effort has been devoted to investigating an appropriate theory for the ductile level. This is not the case for the deformation of metals and there has been considerable effort in trying to understand localization phenomena in metals, which cannot be explained by the standard constitutive theory.

An excellent summary leading to the formulation of the energy theory of localization in metals can be found in two companion papers (Cherukuri and Shawki, 1995a,b). These papers, pointed out to us during the reviewing process, show how engineering developments parallel the recent advances in understanding ductile shear zones in Earth sciences. It is not surprising that the basic conclusions are compatible, thus giving an incentive for future research across the two disciplines. Cherukuri and Shawki postulate that localization phenomena in thermal-elasto-viscoplastic materials can be fully assessed by three independent numbers affecting the energy equation. The first number describes thermal conduction and is the local Peclet number, the second number is the mechanical dissipation or local shear heating number and the third number the local Reynolds number describing the local level of kinetic energy achieved during deformation.

The last dimensionless number defines the biggest difference between ductile Earth- and metal-deformation processes although similar conditions can be recovered in ductile earthquakes as shown in this review. Metals conduct heat very rapidly so that they have to be deformed under high Reynolds numbers to be close to adiabatic conditions. Such near-adiabatic conditions are found to be a necessary ingredient for flow localization in metals (Rogers, 1979). For Earth-like parameters, we, however, advise a different three-dimensional localization space, leaving the Peclet $Pe$ and dissipation numbers $Di$ as important ingredients but adding the damage parameter creating new surface energy instead of the Reynolds number. In the following, we will review the effects of shear heating and conduction only. For this, we rewrite the energy equation in a non-dimensional form where the subscript “0” refers to a reference value of the field quantities, which is chosen to be a value where a homogenous solution applies. In order to use this number as a localization criterion, it is convenient to define a critical dissipation and Peclet number with reference to a homogeneous state just before reaching meta-stability:

$$\frac{\partial T}{\partial t} = Di \sigma_{ij} \dot{e}_{ij} - \frac{1}{Pe} \kappa \nabla^2 T$$  

\hspace{1cm} (A35)

where we neglect the dissipation due to volume changes and the dissipation number is

$$Di = \frac{\psi \sigma_0 \dot{e}_0}{c_p T_0 t_0}$$  

\hspace{1cm} (A36)

The dissipation number $Di$ is the ratio of thermal energy produced by shear heating in the time interval $t_0$ over the energy required to raise the temperature to $T_0$. In terms of a thermodynamically based criterion for departures from the elliptical solutions in extended Gibbs space (Eq. (A21)), we would need to specify both critical Peclet $Pe$ and dissipation numbers $Di$. It turns out that the flow localization phenomenon is only weakly dependent on the Peclet number (Regenauer-Lieb and Yuen, 2003) so that we suggest to characterize ductile materials by a critical dissipation number to describe their tendency to localization. The Peclet number $Pe$ gives the ration of heat transfer by conduction over the heat transfer by advection.

$$Pe = \frac{v_0 \nabla T_0}{\kappa \nabla^2 T_0} = \frac{v_0 L_0}{\kappa}$$  

\hspace{1cm} (A37)

where $v_0$ is the relative velocity of the reference volume (fluid parcel) with respect to a neighbouring
volume and $L_0$ its length scale. Since the post-bifurcational shear zone width is only weakly dependent on the dissipation number but chiefly depends on the Peclet number (Regenauer-Lieb and Yuen, 2003), a good material description can be given by quoting the equilibrium Peclet number for a post-bifurcational steady state solution. The subscript “0” would in this case refer the post-bifurcational thermodynamic equilibrium state. Characterizing the behaviour of complex, composite rheology of the lithosphere in terms of critical dissipation number and equilibrium Peclet number opens the way to incorporate the localization behaviour of the complete lithosphere rheology, obtained from high resolution feedback calculations, into a generic upscaled rheology useful for large scale-coupled mantle convection calculations, which could be a solution to the problems described in the section on applications to plate tectonics.

In engineering sciences, a simple energy theory for localization has been developed much earlier. Shawki (Shawki, 1994a,b) neglects the thermal-elastic feedback term and uses Eq. (A35), solved together with the momentum and the continuity equations, to come up with an energy theory of localization for a single fault in pure shear, i.e. the class of 1D shear zone models discussed in the review and shown schematically in Fig. 2. He solves by linear stability analysis a perturbed initial solution of homogenous simple shear flow. Using the rheologies discussed here, he derives a critical energy criterion and a critical wavelength threshold for growth of perturbations, which also serves as a scaling length for shear zone width.

Shawki's energy criterion for localization is a variance to the one proposed in Eq. (A21). Shawki uses the fact that prior to visible flow localization stationary elastic body waves are emitted, finally guiding visco-plastic bifurcation. This has already been pointed out for the classical constitutive theory in the elasto-plastic case. A good example showing an elastic energy wave preceding elasto-visco-plastic shear zone formation is shown in Fig. 5 of Regenauer-Lieb and Yuen (2000b). While this effect is very difficult to capture numerically, we have suggested to rephrase the criterion into a thermodynamically based approach described by a pair of critical dissipation number and Peclet numbers needed for onset of localization. When simplifying the analytical results of Shawki, the dissipation number can be isolated as the critical number, which is also valid as a global criterion for many interacting faults with more composite complicated lithosphere rheology (Regenauer-Lieb and Yuen, 2003). The results of Shawki and coworkers differs from our afore mentioned papers only concerning the role of elasticity for shear zone nucleation. Shawki concludes that elasticity does not enter the shear zone nucleation criterion but plays an important role on the subsequent shear zone width. This is true if the thermal-elastic feedback term is neglected in the energy equation. We find that with thermal-elastic feedback all localization phenomena require three orders of magnitudes lower dissipation numbers than in comparable cases without thermal-elasticity. Thermal-elasticity acts like a booster to heterogeneous, thermal–mechanical, ductile shear zones.

Of particular scientific concern is the well-posedness of the scientific problem. For this a proof of existence of a unique homogenous solution for the initial boundary conditions must be given. We have mentioned that, in terms of thermodynamics, a positive intrinsic dissipation (Eq. (A21)) is a necessary but not a completely sufficient condition for the existence of a homogenous solution. For the simple shear 1D case, Shawki showed that a unique exact homogenous solution exists for velocity controlled boundaries only if adiabatic (thermally insulating) boundaries are selected. We found equivalent homogenous solutions for the case of a similar pure shear setup (Regenauer-Lieb and Yuen, 2003).

We conclude that Shawki’s energy theory, with the suggested amendments, is a suitable approach for geological materials, if we assume only “simple” feedback between conduction and shear heating. Localization in such simple ductile materials appears to be entirely controlled by the two non-dimensional numbers appearing in the truncated energy Eq. (A35). Considering the low values of diffusivity of rocks, the two numbers have different implications. While the critical dissipation number for transition from homogeneous to bifurcating solutions is giving a criterion for onset of localization the Peclet number is controlling the final width of thermo-mechanical shear zones. These two quantities describe intrinsic thermo-dynamic functions and therefore are constitutive properties, too. But in contrast to the classical brittle theory the ductile theory of localization relies entirely on energy fluxes.
References


Chester, F.M., Higgs, N.G., 1992. Multimechanism friction con-


Green, A.E., Naghdi, P.M., 1965. A dynamical theory of interacting...
Lemonds, J., Needledman, A., 1986. Finite element analyses of shear...


Klaus Regenauer-Lieb is a Research Associate/Lecturer at the Institute of Geophysics ETH Zürich and Privatdozent at the Department of Geosciences of the University of Mainz. He graduated in Geophysics at the University Kiel and obtained a PhD in Geology at the University of Auckland. He held a post-doctoral in the GEOMAR Kiel, in the University of Auckland and at the University of Mainz where he finished in 1999 with a habilitation on energy estimates for large-scale continental deformation. His research interests are in computational geodynamics, with special interest in the coupling of lithosphere and mantle dynamics.
Dave Yuen started out in Physical Chemistry and received his Bachelor’s degree from Caltech. He then switched over to the earth sciences in the aftermath of plate tectonics and received a Master’s degree from Scripps Institution of Oceanography in 1973 and his Doctoral degree under Jerry Schubert at UCLA in 1978. After spending 2 years with Dick Peltier at the Univ. of Toronto, Dave went on to Arizona State University in 1980, then to Univ. Colorado in 1985, and since 1985 he has been at the University of Minnesota at both the Minnesota Supercomputing Institute and the Dept. of Geology and Geophysics. He works on geophysical fluid dynamical problems ranging from the microscales, using molecular dynamics to large-scale circulation problems in mantle convection.