Geophysical Journal International

Geophys. J. Int. (2019) **216,** 365–379
Advance Access publication 2018 October 22
GJI Geodynamics and tectonics

doi: 10.1093/gji/ggy434

Resolving thermomechanical coupling in two and three dimensions: spontaneous strain localization owing to shear heating

T. Duretz [1,2] L. Räss [2,3] Y.Y. Podladchikov [2,3] and S.M. Schmalholz [2]

[1] *Géosciences Rennes, Univ. Rennes* 1, *UMR CNRS* 6118, *35042 Rennes, France. E-mail: thibault.duretz@univ-rennes1.fr*
[2] *Faculté des géosciences et de l'environnement, Institut des Sciences de la Terre, Université de Lausanne, 1015 Lausanne, Switzerland*
[3] *Swiss Geocomputing Centre, Université de Lausanne, 1015 Lausanne, Switzerland*

Accepted 2018 October 18. Received 2018 September 19; in original form 2018 April 20

SUMMARY

Numerous geological processes are governed by thermal and mechanical interactions. In particular, tectonic processes such as ductile strain localization can be induced by the intrinsic coupling that exists between deformation, energy and rheology. To investigate this thermomechanical feedback, we have designed 2-D codes that are based on an implicit finite-difference discretization. The direct-iterative method relies on a classical Newton iteration cycle and requires assembly of sparse matrices, while the pseudo-transient method uses pseudo-time integration and is matrix-free. We show that both methods are able to capture thermomechanical instabilities when applied to model thermally activated shear localization; they exhibit similar temporal evolution and deliver coherent results both in terms of nonlinear accuracy and conservativeness. The pseudo-transient method is an attractive alternative, since it can deliver similar accuracy to a standard direct-iterative method but is based on a much simpler algorithm and enables high-resolution simulations in 3-D. We systematically investigate the dimensionless parameters controlling 2-D shear localization and model shear zone propagation in 3-D using the pseudo-transient method. Code examples based on the pseudo-transient and direct-iterative methods are part of the M2Di routines (Räss et al., 2017) and can be downloaded from Bitbucket and the Swiss Geocomputing Centre website.

Key words: Finite-difference methods; Continental deformation; Numerical techniques; Geodynamics.

1 INTRODUCTION

Thermomechanical feedback represents a first-order multiphysics coupling for geodynamic processes. For instance, thermomechanical coupling plays a major role in initiating and regulating convection currents at the scale of the Earth's mantle (Pekeris 1935; McKenzie *et al.* 1974; Parsons & McKenzie 1978). Thermal convection represents a type of Rayleigh–Bénard instability that is intrinsically linked to the temperature sensitivity of rock-forming mineral densities. Besides convection, thermally activated shear localization is another, yet far less explored, example of thermomechanical coupling in geodynamics. However, these processes have been used to explain the initiation of subduction (Regenauer-Lieb *et al.* 2001; Thielmann & Kaus 2012), the generation of deep earthquakes (Ogawa 1987; Hobbs & Ord 1988; Prieto *et al.* 2013; Ohuchi *et al.* 2017), ductile strain localization (Fleitout & Froidevaux 1980; John *et al.* 2009) or the formation of tectonic nappes (Jaquet & Schmalholz 2017). Thermally activated shear localization occurs when local temperature perturbations owing to shear heating (mechanical dissipation induced by irreversible deformation) are large enough not to be diffused away efficiently. Since ductile mineral strengths strongly depend on temperature (e.g. Carter & Ave'Lallemant 1970), a local temperature increase results in thermal softening, which can further induce the focusing of strain into a localized shear zone (Yuen & Schubert 1979; Fleitout & Froidevaux 1980; Kaus & Podladchikov 2006). Since localization is driven by a self-regulating feedback process, it can either vanish, be maintained stably or evolve into a runaway instability given specific conditions (e.g. Rice & Fairbridge 1975; John *et al.* 2009; Braeck & Podladchikov 2007). Shear zones caused by shear heating have an inherent width (Duretz *et al.* 2014; Moore & Parsons 2015) in which the strain is focused although the thermal imprint may be diffused (Takeuchi & Fialko 2012; Schmalholz & Duretz 2015). Thermally activated shear localization is generally not an exclusive mechanism and can occur in conjunction with other physical processes such as microstructural evolution (Peters *et al.* 2015; Thielmann *et al.* 2015) and mineral reactions (Andersen *et al.* 2008) that can further promote strain localization.

In the following, we first proceed to a dimensional analysis and parameter reduction for thermomechanical equations, which are further used to model thermally activated shear localization. We

introduce two numerical modelling approaches, namely a classic direct-iterative (DI) method and a less conventional, pseudo-transient (PT) approach. The PT method's viability is demonstrated by providing a quantitative comparison with numerical solutions achieved with the standard DI method. We assess both the accuracy and performance of PT solutions. We provide a systematic parametric analysis of 2-D thermally activated shear localization. Natural shear localization always occurs in 3-D, which is important for localization caused by thermal softening, since thermal diffusion of local heat sources is more efficient in 3-D than in 2-D. Yet, there are very few efficient 3-D algorithms that allow one to accurately simulate thermally activated strain localization in 3-D. Here, we show that the PT method is well suited for the efficient and accurate simulation of 3-D shear zone formation by thermal softening. Finally, we discuss how different numerical treatments of nonlinearities and time integration schemes can affect model predictions as well as performance. For reproducibility purposes, we provide the PT and DI numerical codes (MATLAB) used to solve thermomechanical problems. The MATLAB routines are part of M2Di (Räss et al., 2017) and are available for download from Bitbucket at https://bitbucket.org/lraess/m2di and from the Swiss Geocomputing Centre website http://wp.unil.ch/geocomputing/software/. The PT routines are located in the TM2Dpt folder and the DI routines (TM2Di) are located in the M2Di2 folder. The GPU (C-CUDA) routines are available upon request to the authors.

2 THE MATHEMATICAL MODEL

2.1 Thermomechanical coupling

The equations governing thermomechanics of slowly creeping incompressible power-law viscous fluids, in the absence of buoyancy forces, are

$$\begin{aligned}
\frac{\partial v_i}{\partial x_i} &= 0,\\
\frac{\partial \tau_{ij}}{\partial x_j} - \frac{\partial p}{\partial x_i} &= 0,\\
\tau_{ij}\dot{\epsilon}_{ij} + k\frac{\partial^2 T}{\partial x_i^2} - \rho C_\mathrm{p}\frac{\partial T}{\partial t} &= 0,\\
\dot{\epsilon}_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) &= \frac{1}{2}A\tau_\mathrm{II}^{n-1}\exp\left(-\frac{Q}{R\,(T_0+T)}\right)\tau_{ij},
\end{aligned} \tag{1}$$

where v_i are components of the velocity vector in the x_i spatial direction, p is the pressure, T is the temperature deviation from the initial temperature T_0, ρ is the density, C_p is the specific heat, $\dot{\epsilon}_{ij}$ is the strain rate tensor, τ_{ij} and τ_II are the deviatoric stress tensor and the square root of its second invariant:

$$\tau_\mathrm{II} = \sqrt{\frac{1}{2}\tau_{ij}\tau_{ij}}, \tag{2}$$

A is the pre-exponent, n is the stress exponent, Q is the activation energy and R is the universal gas constant. Since we mostly focus on small strains, we do not consider heat transport owing to advection.

Four independent scales

$$\begin{aligned}
\bar{T} &= \frac{nRT_0^2}{Q},\\
\bar{\tau} &= \rho C_\mathrm{p}\bar{T},\\
\bar{t} &= 2^{1-n}A^{-1}\bar{\tau}^{-n}\exp\left(\frac{Q}{RT_0}\right) \text{ and}\\
\bar{L} &= \sqrt{\frac{k}{\rho C_\mathrm{p}}\bar{t}}
\end{aligned} \tag{3}$$

for temperature, stress, time and length, respectively, and their dependent combinations, such as the velocity scale: $\bar{V} = \bar{L}/\bar{t}$, are used to make all the variables dimensionless. Hereinafter, all variables are dimensionless, unless noted otherwise. Introducing dimensionless variables in eq. (1) results in the following dimensionless form of the governing equations (see Appendix A for details):

$$\begin{aligned}
\frac{\partial v_i}{\partial x_i} &= 0,\\
\frac{\partial \tau_{ij}}{\partial x_j} - \frac{\partial p}{\partial x_i} &= 0,\\
\tau_{ij}\dot{\epsilon}_{ij} + \frac{\partial^2 T}{\partial x_i^2} - \frac{\partial T}{\partial t} &= 0,\\
\dot{\epsilon}_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) &= 2^{-n}\tau_\mathrm{II}^{n-1}\exp\left(\frac{n\,T}{1+\frac{T}{T_0}}\right)\tau_{ij},
\end{aligned} \tag{4}$$

The model parameters are the dimensionless initial temperature T_0, the power-law exponent n as well as four parameters that arise from the initial and boundary conditions: the radius and the magnitude of the circular thermal perturbation, the length of the square computational domain, L and the boundary velocity, V_BC (Fig. 2).

3 THE NUMERICAL METHODS

The system of nonlinear equations (eq. 4) is discretized on a Cartesian staggered grid with regular grid-spacing. The time derivative of the heat equation is approximated by either a backward-Euler or a Crank–Nicolson scheme. The dimensionless viscosity is a nonlinear function of both the temperature and strain rate and is expressed, after eq. (4), as

$$\eta = \dot{\epsilon}_\mathrm{II}^{\frac{1-n}{n}}\exp\left(-\frac{T}{1+\frac{T}{T_0}}\right), \tag{5}$$

where $\dot{\epsilon}_\mathrm{II}$ is the square root of the second invariant of the deviatoric strain rate tensor:

$$\dot{\epsilon}_\mathrm{II} = \sqrt{\frac{1}{2}\dot{\epsilon}_{ij}\dot{\epsilon}_{ij}}. \tag{6}$$

Obtaining a numerical solution that satisfies the nonlinear discrete thermomechanical equations is generally an iterative process. Arbitrary initial pressure, velocity and temperature fields will not satisfy the discrete equations as they would produce an imbalance. An implicit solution procedure will seek to iteratively reduce the imbalances until the thermomechanical equations are satisfied to a desired accuracy. To this end, the thermomechanical eq. (4) is

formulated as

$$\begin{aligned} -\frac{\partial v_i}{\partial x_i} &= f_{\mathrm{p}}, \\ \frac{\partial \tau_{ij}}{\partial x_j} - \frac{\partial p}{\partial x_i} &= f_{\mathrm{v}}, \\ \frac{\partial^2 T}{\partial x_i^2} + \tau_{ij}\dot{\epsilon}_{ij} - \frac{\partial T}{\partial t} &= f_{\mathrm{T}}, \end{aligned} \tag{7}$$

where the ∂ symbolises numerical approximation to partial derivatives. The right-hand side terms are nonlinear continuity, momentum and thermal residuals ($f_{\mathrm{p}}, f_{\mathrm{v}}, f_{\mathrm{T}}$), which quantify the magnitude of the imbalance of the thermomechanical equations.

Two methods are employed to minimize the magnitude of theses residuals and deliver accurate pressure, flow and temperature fields. The first method relies on Newton iterations and uses a DI scheme, thus requiring the assembly and factorization of sparse matrices. The second method uses a PT approach, which is fully iterative and matrix-free. Both numerical methods rely on a two-way coupling, since both the coupling term (shear heating) and rheology are treated implicitly.

3.1 The direct-iterative method

For the DI method, we employ a Newton scheme which allows us to obtain accurate nonlinear solutions within few iterations (see Appendix B). The numerical solution, $\mathbf{x} = [\mathbf{v}, \mathbf{p}, \mathbf{T}]^T$, is iteratively corrected with the following update:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha \delta \mathbf{x}^{k+1}, \tag{8}$$

where the δ operator stands for the correction of a quantity, α is a scalar optimization parameter and k is the nonlinear iteration index. The Newton correction, $\delta\mathbf{x} = [\delta\mathbf{v}, \delta\mathbf{p}, \delta\mathbf{T}]^T$, is obtained by applying the inverse of the Jacobian matrix, $\mathbf{J}_{\mathrm{TM}}$, to the current nonlinear residual $\mathbf{f} = \left[\mathbf{f}_{\mathrm{v}}, \mathbf{f}_{\mathrm{p}}, \mathbf{f}_{\mathrm{T}}\right]^T$:

$$\delta \mathbf{x}^{k+1} = -\ {\mathbf{J}_{\mathrm{TM}}}^{-1}\ \mathbf{f}^k. \tag{9}$$

Prior to the solution update, we run a line search procedure to determine the optimization parameter α ($0 < \alpha \leq 1$) that yields:

$$\min \left|\left|\mathbf{f}\left(\mathbf{x}^k + \alpha\ \delta \mathbf{x}^{k+1}\right)\right|\right|_{\mathrm{L2}}. \tag{10}$$

In the DI context, it is necessary to formulate and assemble the Jacobian matrix. The latter describes the gradient of the residuals with regard to the solutions; for example, the sensitivity of the momentum imbalance with regard to the velocity field. For the considered thermomechanically coupled flow, the Jacobian matrix takes the form of

$$\mathbf{J}_{\mathrm{TM}} = \frac{\partial \mathbf{f}_i}{\partial \mathbf{x}_j} = \begin{bmatrix} \mathbf{J}_{\mathrm{vv}} & \mathbf{K}_{\mathrm{vp}} & \mathbf{J}_{\mathrm{vT}} \\ \mathbf{K}_{\mathrm{pv}} & \mathbf{0} & \mathbf{0} \\ \mathbf{J}_{\mathrm{Tv}} & \mathbf{0} & \mathbf{J}_{\mathrm{TT}} \end{bmatrix}. \tag{11}$$

The $\mathbf{J}_{\mathrm{vv}}$ corresponds to the mechanical Jacobian matrix, which arises from the strain rate dependence of viscosity. The $\mathbf{J}_{\mathrm{vT}}$ block highlights the temperature dependence of viscosity and the $\mathbf{J}_{\mathrm{Tv}}$ block arises from the linearization of the shear heating term. The $\mathbf{K}_{\mathrm{vp}}$ and $\mathbf{K}_{\mathrm{pv}}$, respectively, represent the gradient operator and divergence discrete operators. The matrix $\mathbf{J}_{\mathrm{TT}}$ is a modified Laplace operator that also includes contributions from the temperature dependence of the viscosity. The Newton corrections for the thermomechanical system can be formulated as

$$\delta \mathbf{x} = \begin{bmatrix} \delta\mathbf{v} \\ \delta\mathbf{p} \\ \delta\mathbf{T} \end{bmatrix} = -\begin{bmatrix} \mathbf{J}_{\mathrm{vv}} & \mathbf{K}_{\mathrm{vp}} & \mathbf{J}_{\mathrm{vT}} \\ \mathbf{K}_{\mathrm{pv}} & \mathbf{0} & \mathbf{0} \\ \mathbf{J}_{\mathrm{Tv}} & \mathbf{0} & \mathbf{J}_{\mathrm{TT}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_{\mathrm{v}} \\ \mathbf{f}_{\mathrm{p}} \\ \mathbf{f}_{\mathrm{T}} \end{bmatrix}, \tag{12}$$

and is obtained via DI procedure (see Appendix C).

The nonlinear iteration cycle is aborted once one of the following criteria $||\mathbf{f}||_{\mathrm{L2}} / ||\mathbf{f}||_{\mathrm{L2}}^{\mathrm{initial}} < \mathrm{tol}_{\mathrm{nonlin}}^{\mathrm{rel}}$ or $||\mathbf{f}||_{\mathrm{L2}} < \mathrm{tol}_{\mathrm{nonlin}}^{\mathrm{abs}}$ is met; meaning that the thermomechanical balance equations are satisfied to the desired accuracy. The main steps of the DI approach are summarized in Fig. 1(a). This code is written in MATLAB language; it is based on and made available as part of the M2Di routines (Räss *et al.* 2017) under the name TM2Di (https://bitbucket.org/lraess/m2di).

3.2 The pseudo-transient method

The PT or relaxation method allows one to iteratively solve nonlinear problems in a single iteration loop in a matrix-free way. The relaxation method is a classical numerical technique to solve stationary (elliptic) problems (Frankel 1950). The method was extended in the 1960s to elastic problems (Otter *et al.* 1966) and more recently to elasto-plastic (Cundall 1982) and viscoelastic problems (Poliakov *et al.* 1993). The PT method relies on introducing PT terms into steady-state equations. Given a set of initial and boundary conditions, solutions can be found by integrating the equations forward in pseudo-time (τ) until steady state is attained; that is, the pseudo-time derivative vanishes. For example, the compressible Navier–Stokes equations incorporate right-hand side time derivatives for both the mass $\left(\beta \frac{\partial p}{\partial t}\right)$ and momentum $\left(\rho \frac{\partial v_i}{\partial t}\right)$ balance equations. These latter represent the elastic bulk rheology (with β as compressibility) and the acceleration (with ρ as the density), respectively. A solution to the incompressible Stokes problem requires that both of these transient terms vanish.

The essence of the PT method is to integrate the balance equations in pseudo-time, τ, until the PT terms vanish. To this end, eq. (4) is expressed as

$$\begin{aligned} -\frac{\partial v_i}{\partial x_i} &= \beta \frac{\partial p}{\partial \tau_{\mathrm{p}}}, \\ \frac{\partial \tau_{ij}}{\partial x_j} - \frac{\partial p}{\partial x_i} &= \rho \frac{\partial v_i}{\partial \tau_{\mathrm{v}}}, \\ \frac{\partial^2 T}{\partial x_i^2} + \tau_{ij}\dot{\epsilon}_{ij} - \frac{\partial T}{\partial t} &= \frac{\partial T}{\partial \tau_{\mathrm{T}}}. \end{aligned} \tag{13}$$

This approach is equivalent to iteratively reducing the magnitude of the residual of each equation (e.g. as in the DI method) since PT terms are equivalent to residuals.

The nonlinear viscosity η is evaluated at each PT iteration k using the current strain rate and temperature solution fields. The treatment of nonlinearities is greatly facilitated by using an effective viscosity (η_{eff}), which we formulate as

$${\eta_{\mathrm{eff}}}^k = \exp\left[\theta_\eta \ln\left({\eta_{\mathrm{eff}}}^{k-1}\right) + (1-\theta_\eta)\ln\left(\eta^k\right)\right], \tag{14}$$

where θ_η ($0 \leq \theta_\eta \leq 1$) corresponds to a relaxation factor. This approach is a continuation method, since the effective viscosity progressively relaxes towards the de facto physical viscosity ($\eta_{\mathrm{eff}} \rightarrow \eta$) throughout the PT iterations.

The integration of the momentum, mass conservation and temperature evolution equations necessitates the definition of individual pseudo-time steps, $\Delta\tau_{\mathrm{v}_i}$, $\Delta\tau_{\mathrm{p}}$ and $\Delta\tau_{\mathrm{T}}$. Hereinafter, we assume that ρ and β are equal to 1.0 and the pseudo-time steps are formulated

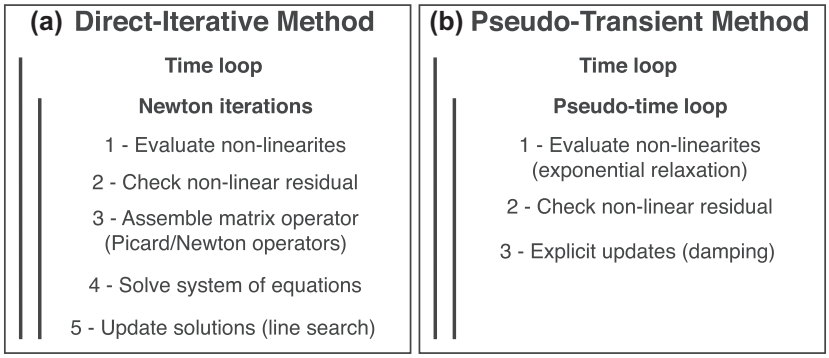

Figure 1. Algorithmic flowchart for both methods used in the study: (a) the direct-iterative method (TM2Di code) and (b) the pseudo-transient method.

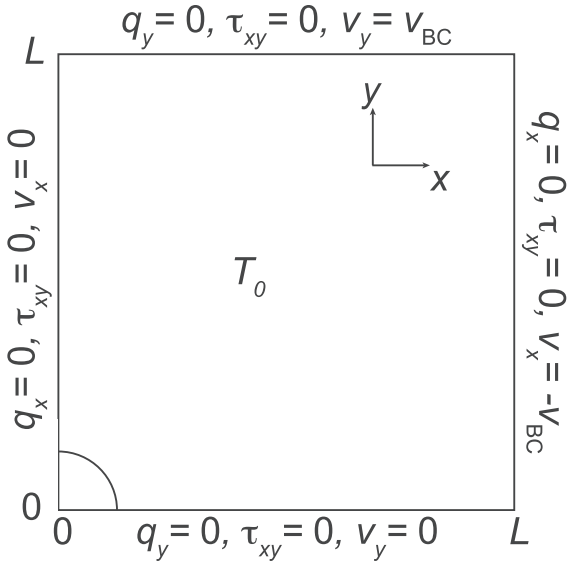

Figure 2. Schematic initial model configuration for 2-D calculations. T_0 stands for the initial temperature, L is the box length and V_{BC} is the boundary velocity.

as

$$\begin{aligned}
\Delta\tau_\mathrm{p} &= \theta_\mathrm{p}\,\frac{2.1\,N_\mathrm{dim}\,\eta_{it}^k\,(1+\eta_\mathrm{b})}{\max(N_x,\,N_y)},\\
\Delta\tau_{\mathrm{v}_i} &= \theta_{\mathrm{v}_i}\,\frac{\min(\Delta x,\,\Delta y)^2}{2.1\,N_\mathrm{dim}\,\eta_{it}^k\,(1+\eta_\mathrm{b})},\\
\Delta\tau_\mathrm{T} &= \theta_\mathrm{T}\frac{\min(\Delta x,\,\Delta y)^2}{2.1\,N_\mathrm{dim}},
\end{aligned} \tag{15}$$

where η_b is a numerical analogy of bulk viscosity, N_x and N_y are the number of gridpoints in the x and y direction, respectively, Δx and Δy are grid-spacings, θ_{v_i}, θ_p and θ_T are pseudo-time step reduction factors ($\leq$1.0) and N_dim is the number of dimensions. $\Delta\tau_\mathrm{T}$ is the pseudo-time step used for explicit integration of diffusion equation satisfying the Courant–Friedrichs–Lewy condition (CFL). $\Delta\tau_{\mathrm{v}_i}$ is the pseudo-time step used for integrating the momentum equations. It is constructed by multiplying the explicit CFL time step for viscous flow $\min(\Delta x,\,\Delta y)^2/(2.1\,N_\mathrm{dim}\,\eta_{it}^k)$ by $1/(1+\eta_\mathrm{b})$, which includes a numerical analogy to the bulk viscosity η_b. The use of the denominator of $\Delta\tau_{\mathrm{v}_i}$ as nominator in the definition of $\Delta\tau_\mathrm{p}$ leads to an empirically derived pseudo-time step for the continuity equation. This criteria allows for an optimal convergence of the Stokes problem, making the iteration strategy less sensitive to the physical shear viscosity η. A dimensional analysis confirms that the product of $\beta^{-1}\Delta\tau_\mathrm{p}$ [Pa s] by the divergence of velocities [s^{-1}] produces dynamic pressure increments in [Pa]. Identical reasoning can be applied to the momentum balance equation, where $\rho^{-1}\Delta\tau_\mathrm{v}$ [$\mathrm{m}^2\,\mathrm{Pa}^{-1}\,\mathrm{m}^{-1}$] multiplies the force balance terms [$\mathrm{Pa\,m}^{-1}$] to produce increments of velocity [$\mathrm{m\,s}^{-1}$]. We further highlight that η_{it}^k refers to entire fields (defined for every gridpoint); thus, pseudo-time step values are local to every gridpoint within the computational domain and analogous to the application of diagonal preconditioner in matrix-based solvers. At each PT iteration, the velocity, pressure and temperature fields are updated at each iteration using their current values of pseudo-time step and residual:

$$\begin{aligned}
p^k &= p^{k-1} + \Delta\tau_\mathrm{p}\,f_\mathrm{p}^k\\
v_i^k &= v_i^{k-1} + \Delta\tau_{\mathrm{v}_i}\,g_{\mathrm{v}_i}^k\\
T^k &= T^{k-1} + \Delta\tau_\mathrm{T}\,f_\mathrm{T}^k.
\end{aligned} \tag{16}$$

The use of damping greatly reduces the number of iterations needed for convergence of the PT iterations (Choi *et al.* 2013; Yang & Mittal 2014). To this end, the damped momentum residual $g_{\mathrm{v}_i}^k$ in eq. (16) is written as

$$g_{\mathrm{v}_i}^k = f_{\mathrm{v}_i}^k + \left(1 - \frac{\nu}{N_i}\right) f_{\mathrm{v}_i}^{k-1} \tag{17}$$

where optimal values of ν reside within the range ($1 \leq \nu \leq 10$) and N_i is the number of gridpoints in the direction i. An analogue approach for elastic rheology is described by Cundall & Strack (1979) and is successfully used in the FLAC geotechnical software (Cundall *et al.* 1993). As for the DI method, the PT iterations are performed until one of the following criteria $||\mathbf{f}||_{\mathrm{L2}} / ||\mathbf{f}||_{\mathrm{L2}}^{\mathrm{initial}} < \mathrm{tol}_{\mathrm{nonlin}}^{\mathrm{rel}}$ or $||\mathbf{f}||_{\mathrm{L2}} < \mathrm{tol}_{\mathrm{nonlin}}^{\mathrm{abs}}$ is verified.

3.3 Physical time integration

For both the DI and PT methods, the integration of the heat equation is done in physical time t. An implicit (backward-Euler) or semi-implicit (Crank–Nicolson) solution is obtained by updating the heat fluxes and shear heating term at each nonlinear or PT iteration. In the following examples, we use a physical time step, Δt_T, which is proportional to the CFL time step:

$$\Delta t_\mathrm{T} = \xi\,\frac{\min(\Delta x,\,\Delta y)^2}{2.1\,N_\mathrm{dim}} = \xi\,\Delta t_\mathrm{exp}, \tag{18}$$

where ξ corresponds to a time step ratio, $\Delta t_\mathrm{T}/\Delta t_\mathrm{exp}$. Despite the use of an implicit scheme, we did not obtain successful time integration of the nonlinear equations system for an arbitrary choice of time

step. Thus, we allow for time step values that are proportional to those required for explicit integration of the heat equation (Δt_{exp}). In the following section, we systematically investigate for different spatial resolution which values of time step can lead to successful time integration. Implementations of the thermomechanical DI and PT solvers were performed in both MATLAB and C-CUDA languages (only for the PT method).

4 A COMPARISON OF THE DIRECT-ITERATIVE AND THE PSEUDO-TRANSIENT METHODS

In this section, we demonstrate that both the DI method and the PT method can solve thermomechanical problems to the same accuracy level. We evaluate each method's performance on current personal computers (single CPU and single GPU).

4.1 The reference model's configuration

The numerical models were designed to study the propagation of shear zones owing to shear heating in a viscous medium subjected to far-field pure shear kinematics. For 2-D calculations, we consider a physical domain of dimensions $[0, L] \times [0, L]$. The normal boundary velocities are set to $v_x = -V_{\text{BC}}$ for $x = L$, $v_y = V_{\text{BC}}$ for $y = L$ and 0 elsewhere. Zero shear stress and zero heat flux boundary conditions are applied on all model sides. The initial temperature field, T_0, is set to 16.4423 and a perturbation of radius equals to $0.0857L$ and amplitude equals to $0.1T_0$ is centred around the origin (Fig. 2).

The reference model is run for a boundary velocity of 66.4437, a model length of 0.86038 and a stress exponent of 3. For the 3-D models, the model domain extends up to $z = L$, where a zero normal boundary velocity, shear stress and heat flux are applied. Thus, the first model embeds an initial cylindrical (i.e. continuous in the z direction) temperature perturbation and is similar to the 2-D model. In the second model, an initially spherical temperature perturbation is prescribed.

4.2 The temperatures and strain rates inside the shear zones

Thermally activated shear zones have inherent length scales that are proportional to material properties and loading conditions. Thus, it is possible to characterize dynamically evolving fields such as the temperature in such shear zones. If the numerical resolution is greater than the characteristic length scale, these modelling results are essentially independent of the numerical resolution (e.g. Duretz *et al.* 2014). A typical model evolution is depicted in Fig. 3, which shows the progressive focusing of strain rate and temperature with time. Since the shear bands have a finite-length scale, it is possible to monitor the evolution of these quantities inside the shear band (Fig. 4). The strain rate overcomes the background strain rate by an order of magnitude within 1.5×10^{-3} time units. The temperature increase follows a distinct evolution and tends towards a value of 3.5 for a model time of 3.0×10^{-3}. The simulations computed with the DI and PT methods clearly provide the same temperature and strain rate predictions (Fig. 4).

4.3 The nonlinear solvers' accuracy

To show the PT method's ability to handle nonlinear thermomechanical problems, we present a quantitative analysis of errors caused by the nonlinearity. A single time step numerical solution was computed with the DI method using the previously described setup and a resolution of 94^2 numerical cells. The solution procedure was aborted once we attained machine precision for nonlinear residuals; the obtained effective strain rate, temperature and pressure serve as reference solution fields. We then computed a series of numerical solutions using larger nonlinear tolerances (iteration exit criteria) with both the PT and DI method.

The deviation of the numerical solutions with regard to the reference fields was calculated as

$$\|E_a\|_{L2} = \|a^{\text{numerical}} - a^{\text{reference}}\|_{L2}, \tag{19}$$

where a is either $\dot{\epsilon}_{\text{II}}$, P or T. For the same nonlinear tolerance, the DI and PT provide the same deviation from the reference results (Fig. 5). The deviations obtained with either the DI or the PT decrease as the nonlinear tolerance is decreased. Thus, both methods converge towards the reference numerical solution with a linear trend. Nonlinear numerical solutions obtained with the PT method can reach the same accuracy level as those obtained with the DI method.

4.4 The conservation of energy

For thermomechanical problems, considering a purely viscous rheology, energy conservation postulates that mechanical work should be fully converted into heat. The numerical simulations' accuracy depends on numerical schemes' ability to conserve energy and therefore fulfil the energy conservation equation. The work per unit length is obtained by evaluating the following integrals (Green's theorem):

$$W = \int_t \int_V \tau_{ij} \dot{\epsilon}_{ij} \, \mathrm{d}V \mathrm{d}t = \int_t \oint_S \sigma_{ij} v_i n_j \, \mathrm{d}S \mathrm{d}t, \tag{20}$$

where n_j is the unit vector to the boundaries and S is the surface of the domain. The thermal energy per unit length takes the form of

$$E = \int_t \int_V \frac{\partial T}{\partial t} \, \mathrm{d}V \mathrm{d}t, \tag{21}$$

where V is the volume of the domain. The time evolution of W and E for 2-D numerical simulations using the configuration described in Section 4.1 is depicted in Fig. 6. The work and thermal energy computed from the PT and DI simulations all follow the same trend. For either methods, the work equals to thermal energy at any moment in time. Thus, the numerical solutions arising from the finite-difference discretization are conservative, independent of the solving procedure.

4.5 Performance

We evaluate the performance of the MATLAB CPU-based and C-CUDA GPU-based PT solvers using two different metrics. The first metric is an evaluation of the effective memory throughput ($\text{MTP}_{\text{effective}}$). The second metric is a measurement of the wall-time needed to achieve convergence over a time step. Since the PT solvers perform stencil operations in a matrix-free approach, the memory transfers bound the algorithm and the number of floating-point operations per second are not affecting the performance. The used $\text{MTP}_{\text{effective}}$ metric (Omlin 2017) evaluates how efficiently data is transferred between the memory and the computation units, in Gigabytes per second (GB s^{-1}):

$$\text{MTP}_{\text{effective}} = \frac{(N_x \times N_y) \times N_{\text{t}} \times \text{nIO} \times \text{precis}}{1e9 \times \text{time}_{N_{\text{t}}}}, \tag{22}$$

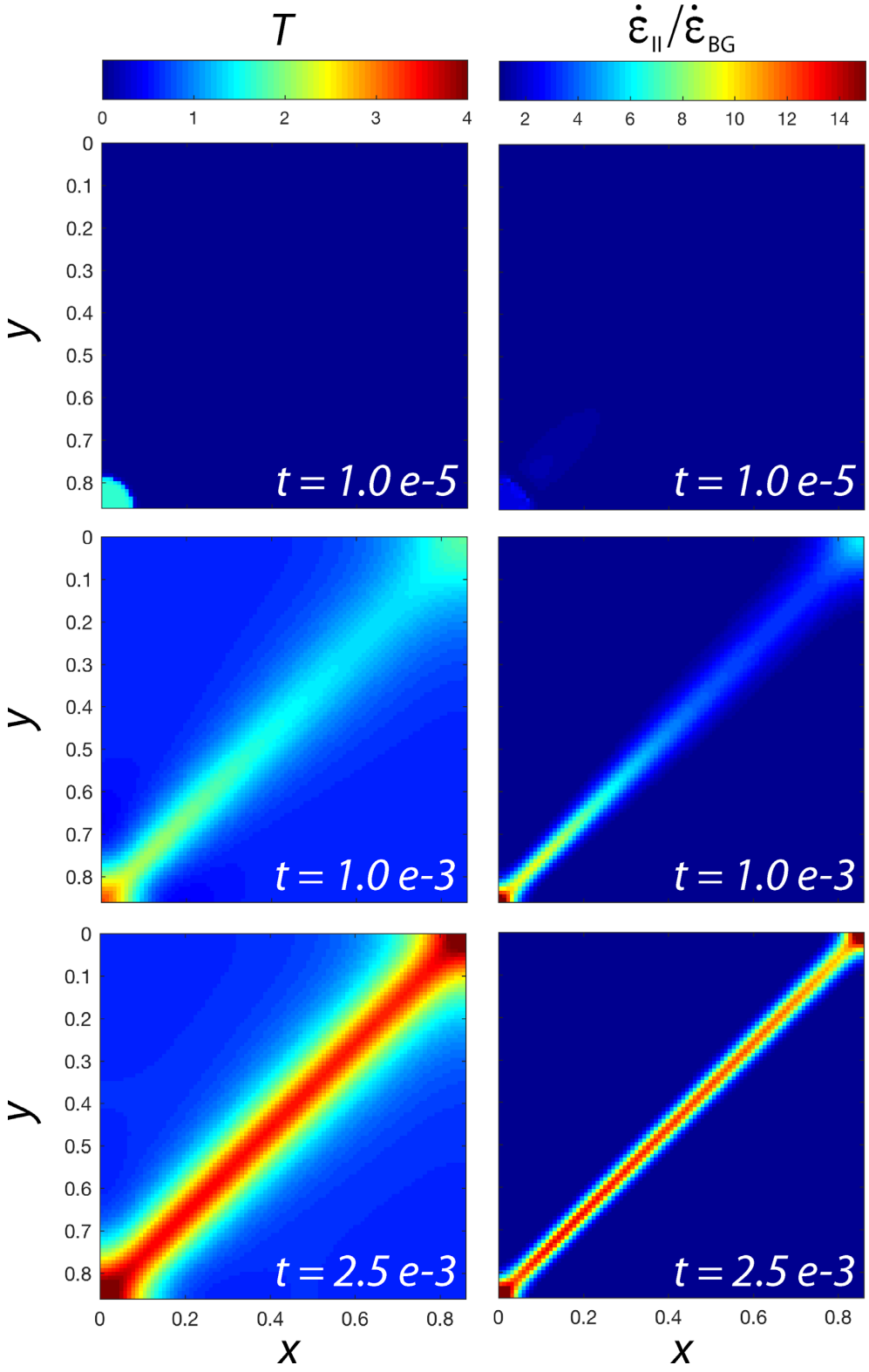

Figure 3. Thermomechanical activated shear localization in 2-D. The model was run using the reference model parameter (see the main text) and initiated with a circular temperature perturbation (10 per cent increase of temperature). Pure shear was applied on boundaries normal to x (inflow) and y (outflow). All thermal boundary conditions were insulating (zero heat flow).

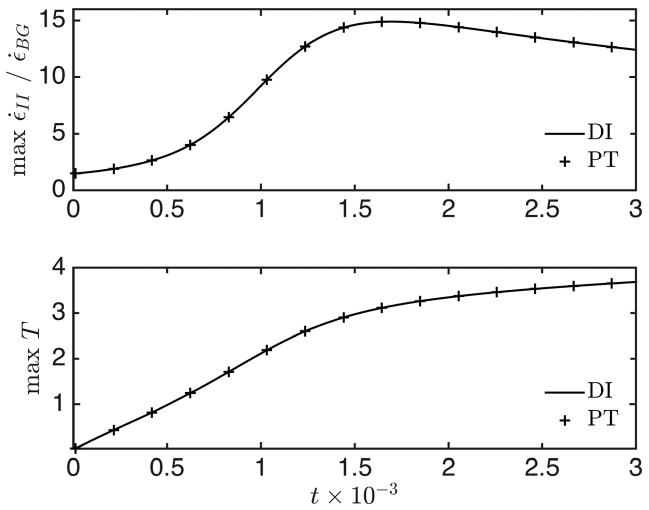

Figure 4. Strain rate amplification (upper panel) and temperature (lower panel) evolution in the shear zone (away from the model boundaries). Both PT and DI models were run with a relative tolerance of 10^{-9} and with backward-Euler time integration for the energy equation. For this example, PT models typically converge within about 10^5 iterations.

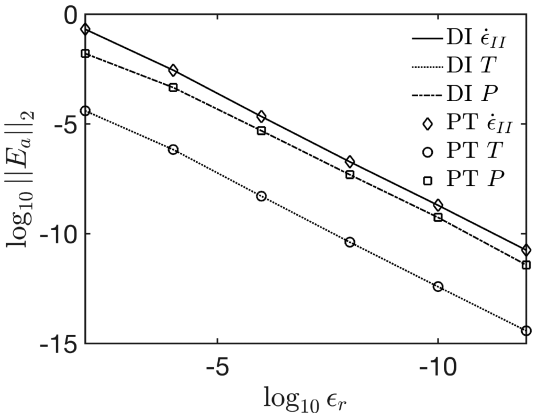

Figure 5. Nonlinear accuracy PT and the DI method (TM2Di code). Errors in P, T and $\dot{\epsilon}_{\rm II}$ were computed for different relative tolerance ($\varepsilon_{\rm r}$). The PT and the DI methods provide similar errors, and both converge linearly to the reference results with decreasing tolerance.

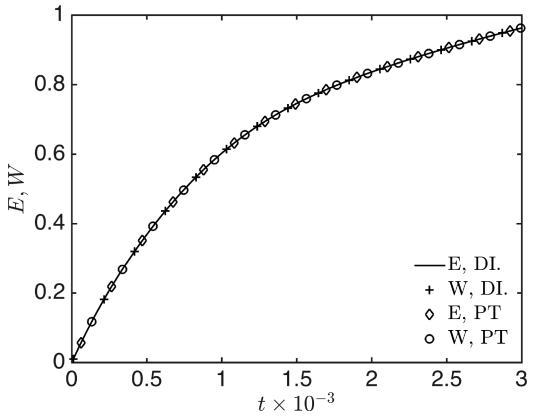

Figure 6. Evolution of mechanical work (W) and heat (E) in time. Both the PT and DI models were run with backward-Euler time integration for the energy equation.

where ($N_x \times N_y$) is the number of cells, $N_{\rm t}$ is the number of time steps or iterations performed, nIO is the number of memory accesses performed, precis is the floating-point precision (either 4 or 8 bytes) and $\text{time}_{N_{\rm t}}$ is the time (in seconds) needed to perform the $N_{\rm t}$ steps. The number of memory accesses (nIO) defines the minimum number of read-and-write or read-only operations required to solve the specific physics. For 2-D coupled thermomechanics, the read-and-write operations correspond to the updates of the degrees of freedom (DoF: v_x, v_y, P, T), and two additional read-only operations for converging the nonlinear viscosity; in our case, nIO = 10. All the performance benchmark runs are performed using double-precision floating-point arithmetic (precis = 8 bytes) for fair comparison in particular between MATLAB and C-CUDA implementations. The $\text{MTP}_{\text{effective}}$ values reported in Fig. 7(a) represent the efficiency of memory transfers for the PT solvers, for both vectorized MATLAB CPU and C-CUDA single-GPU implementations. The obtained numbers should be compared to the peak memory throughput values (MTP_{peak}) for the specific hardware, here an Intel i5 CPU and an Nvidia Titan X (Maxwell) GPU. MTP_{peak} values are measured performing memory copy only without any computation. Values of MTP_{peak} are in the order of 20 GB s^{-1} for the Intel i5 CPU and in the order of 260 GB s^{-1} for the Titan X (Maxwell) GPU. The MATLAB CPU implementation runs at about 2 per cent of the MTP_{peak}, while the C-CUDA GPU codes run above 10 per cent of the MTP_{peak}. The optimized memory bandwidth as well as the inherent parallelism on the GPU chip could explain such differences. The resulting overall performance gain of the parallel GPU implementation versus the serial CPU is more than two orders of magnitude. The GPU $\text{MTP}_{\text{effective}}$ values show that some optimization steps could still be performed in order to achieve closer to MTP_{peak} values. Such considerations are beyond the scope of this study, but could include increased number of on-the-fly computation, kernel rearranging and register queues.

Although the $\text{MTP}_{\text{effective}}$ provides the efficiency of hardware utilization for a specific implementation of the thermomechanical solver, a more relevant metric should be used to compare the memory-bounded stencil PT iterative approach to the DI solver TM2Di. Here, the wall-time metric is chosen to assess the overall time to solution of a nonlinear step converged to $\text{tol}^{\text{rel}}_{\text{nonlin}} = 10^{-8}$ (Fig. 7b). The Newton-based DI solver TM2Di shows a close to linear increase of wall-time with increasing problem size (DoF). It is implemented in MATLAB and an Intel i5 (2016) CPU on a system equipped with 16 GB of RAM is used for computations. The maximal 2-D problem size fitting in 16 GB RAM represents a numerical domain of 960^2 gridpoints, solved in a wall-time close to 2.5 min. In comparison, 15 per cent less time was needed to converge the same problem using the C-CUDA GPU-based PT solver on an Nvidia Titan X (Maxwell). Nonetheless, the key benefit of this method is the maximal problem size that can be resolved using the available 12 GB of on-chip RAM of the GPU; 163 MDoF represents a numerical 2-D domain size of 6400^2 cells, which is a very high numerical resolution compared to the resolution currently used in geodynamic numerical simulations. In terms of wall-time, the PT GPU solver outperforms the Newton DI-based solver TM2Di for the investigated setup.

4.6 Explicit and implicit coupling strategies

The numerical solution of multiphysics problems can be achieved by means of various coupling strategies. For thermomechanical flow,

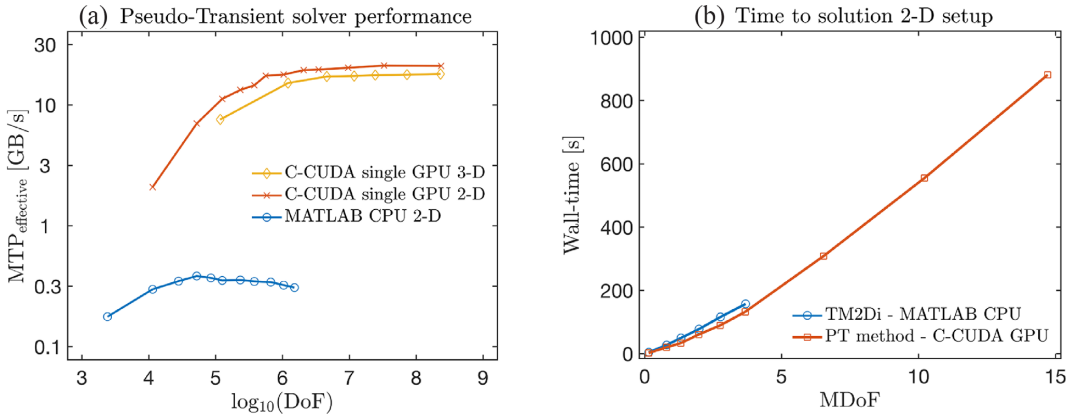

Figure 7. Performance evaluation of the thermomechanically coupled solvers. (a) Effective memory throughput $MTP_{effective}$ in GB s^{-1} of the pseudo-transient implementations using an iterative and matrix-free approach. 2-D MATLAB running on an Intel i5 (2016) processor with 16 GB RAM is compared to 2-D and 3-D C-CUDA running on Nvidia Titan X (Maxwell) GPU. The DoF represents four variables in 2-D (v_x, v_y, P and T) and five variables in 3-D (v_x, v_y, v_z, P and T), multiplied by the respective number of gridpoints. The $MTP_{effective}$ of the 2-D and 3-D GPU implementations saturate at about 20 GB s^{-1}. This is one order of magnitude lower than the MTP_{peak} (memory copy only) measured on the Titan X (Maxwell) GPU. The vectorized MATLAB implementation running on the Intel i5 CPU shows a close to two orders of magnitude discrepancy between peak and effective MTP. Two orders of magnitude of $MTP_{effective}$ is observed between the GPU implementation and that of the CPU, both performing double-precision arithmetics. (b) Time-to-solution converging a nonlinear step to $tol_{nonlin} = 1e-8$, comparing the pseudo-transient method implemented in C-CUDA running on a single Nvidia Titan X (Maxwell) GPU to the direct solver type using the TM2Di Newton MATLAB implementation executed on an Intel i5 (2016) CPU. 3.7 MDoF represents a 2-D domain of 960 x 960 gridpoints and is the maximal resolution that TM2Di can handle while using less than 16 GB of RAM. 2-D domain resolution up to 1920^2 gridpoints could be solved in 15 min using the matrix-free pseudo-transient GPU solver while using less than 12 GB of RAM (on the device).

a two-way coupling implies an implicit treatment of nonlinear coupling terms (Popov & Sobolev 2008; Kaus *et al.* 2016), namely the viscous-dissipation term and the strain rate and temperature dependence of the viscosity. In the two-way coupling, the viscosity is thus a function of the strain rate and temperature evaluated at new time index: $\eta\left(\dot{\epsilon}_{\mathrm{II}}{}^{t+\Delta t}, T^{t+\Delta t}\right)$. One-way coupling represents an alternative coupling strategy commonly used in geodynamic modelling. One-way coupling resides in (1) obtaining a nonlinear solution of the purely mechanical problem (thus only considering the strain rate dependence of viscosity) and (2) solving the energy equation using the mechanical dissipation obtained from the mechanical solution (Kaus & Podladchikov 2006). In other words, one-way coupling uses strain rates at the new time index but temperature of the old time index to evaluate the viscosity, $\eta\left(\dot{\epsilon}_{\mathrm{II}}{}^{t+\Delta t}, T^{t}\right)$. A fully explicit coupling strategy can also be envisaged (Gerya & Yuen 2003, 2007). This approach relies on an explicit treatment of both coupling terms and rheological equations, which results in a linear mechanical problem. With this approach, the viscosity is evaluated using solutions from the previous time index: $\eta\left(\dot{\epsilon}_{\mathrm{II}}{}^{t}, T^{t}\right)$. To evaluate the impacts of the different coupling strategies, we have run our reference simulation (Section 4.1) with the different strategies. Strain localization can be achieved with an explicit coupling strategy, but with the least intensity. We have monitored the gains in strain rate amplification and temperature using the solutions obtained using explicit coupling as reference (Fig. 8). For the same value of physical time step, Δt, one-way coupling leads to an increase of 25 per cent in strain rate amplification and two-way coupling results in a 40 per cent increase in strain rate amplification. The impacts on maximum temperatures were less pronounced, since increments of 8 and 13 per cent were obtained for one-way and two-way coupling, respectively. Refining the time step can be used to improve the accuracy of the one-way coupling approach. For instance, twice smaller time step led to a 30 per cent gain in strain rate; however, with a smaller time step, the gain in strain rate rapidly saturates (here, to about 32 per cent) and does not catch up the values obtained with a two-way coupling. Alternative coupling strategies, such as time averaging whereby $\eta\left(\dot{\epsilon}_{\mathrm{II}}{}^{t+\frac{\Delta t}{2}}, T^{t+\frac{\Delta t}{2}}\right)$, were not considered here but could also be envisaged.

4.7 Time integration

The choice of a time integration scheme is also critical when solving coupled nonlinear transient equations. Using our configuration (Section 4.1), successful time integrations were not possible for arbitrary large time step values. Time step values were often required to be on the order of the CFL criteria for diffusion despite the use of an implicit integration scheme (backward-Euler). This was especially true when using low grid resolutions (50^2 cells, $h = 1.7 \times 10^{-2}$). We have reported the range of time step variation factors ξ that provided stable integration for various grid resolutions (Fig. 9). Restrictions on time step values usually decrease with increasing grid resolution. At high resolution (500^2 cells, $h = 1.7 \times 10^{-3}$), a time step variation factor in the order of 50 was achievable using backward-Euler. The Crank–Nicolson scheme generally proved to have a larger stability domain, and time step variation factors could reach 80 at high resolution. In practice, the combination of a Crank–Nicolson scheme with an adaptive time stepping procedure (e.g. using time step bisection based on the magnitude of nonlinear residuals (Popov & Sobolev 2008; Duretz *et al.* 2015) can allow for stable and flexible time integration.

5 THERMALLY ACTIVATED SHEAR LOCALIZATION: APPLICATIONS

Here, we present applications computed with the PT approach that demonstrate the method's flexibility as well as its power.

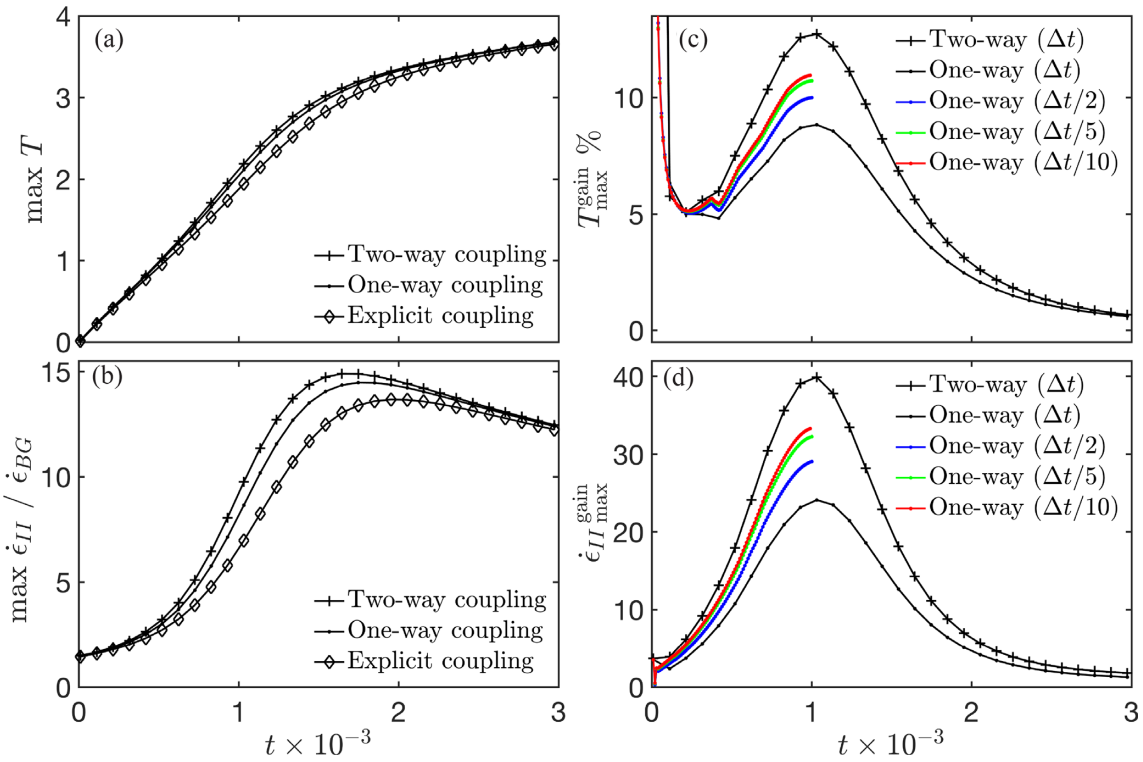

Figure 8. The impacts of different multiphysics coupling strategies (one-way coupling, two-way coupling and explicit coupling) on numerical solution. Panels (a) and (b) depict the evolution of the temperature increase and strain rate amplification in the shear zone for the different coupling strategies using the same constant physical time step (Δt). Panels (c) and (d) show the gains in temperature and strain rate when using one-way coupling and two-way coupling in comparison to the explicit coupling (used as a reference here). We also report results obtained with the one-way strategy but with lower time step values (colour lines). Backward-Euler time integration was used to integrate the energy equation.

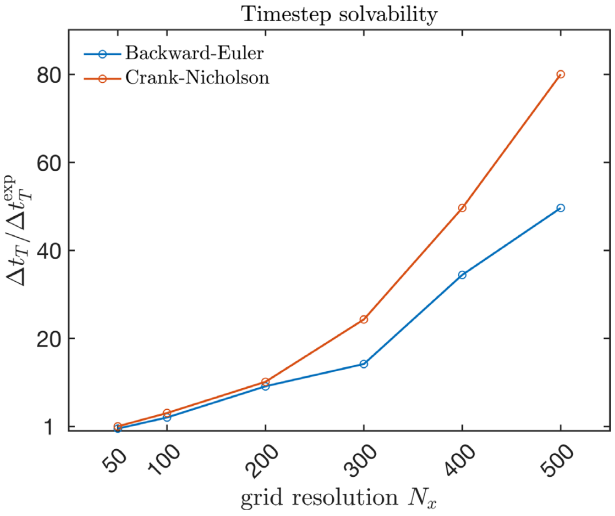

Figure 9. Solvability of the thermomechanical nonlinear system (two-way coupling). Models were run for different values of $\xi = \Delta t_{\text{T}}/\Delta t_{\text{T}}^{\text{exp}}$ and various grid resolutions N_x. At larger resolutions, models with a larger physical time step become more solvable. These results were obtained using the DI method. Models were run for the reference parameters until a final time of 3×10^{-3}. Simulations for which the DI solver failed to converge for the requested value of ξ are considered unsuccessful. In a general case, the time step could be adapted through the simulations in case the linear or nonlinear solves are unsuccessful.

5.1 Thermomechanical strain localization in 2-D: a systematic study

We have studied the relative importance of the boundary velocity (V_{BC}) and bulk strain (ε) on shear zone development and evolution, performing 139 systematic 2-D simulations. Each simulation completed 1000 time steps with a resolution of 190^2 cells. We have used a Crank–Nicolson time integration and relative nonlinear tolerance of 10^{-5}. The entire systematic study was run sequentially on a single Nvidia GTX Titan X (Maxwell) GPU card. We have monitored the maximum temperature and the strain rate amplification factor in the evolving shear zone (Figs 10a and b) using the model configuration described above (Section 4.1). Strain localization occurred over the entire parameter space to a variable degree. Both the maximum temperature and strain rate amplification strongly depend on V_{BC} and ε. Weak shear localization occurred for low-boundary velocity and is characterized by a small strain rate amplification factor (<5). The most intense shear localization led to peak temperature (>8) and strain rate amplification factor($>10^2$) and was achieved for a boundary velocity of 400.

In-depth analysis of this data set reveals that the model results can be collapsed using a single parameter $V_{\text{BC}}\log_{10}\varepsilon$. Variations of this non-dimensional parameter allow one to predict the maximum temperature and strain rate amplification over the entire investigated parameter range (Figs 10c and d). The consistent collapse for simulations with different bulk strains and significantly variable localization intensities (different temperatures and strain rates in the shear zone) further show the approach's robustness, since the accuracy of numerical solutions is the same over the investigated parameter range.

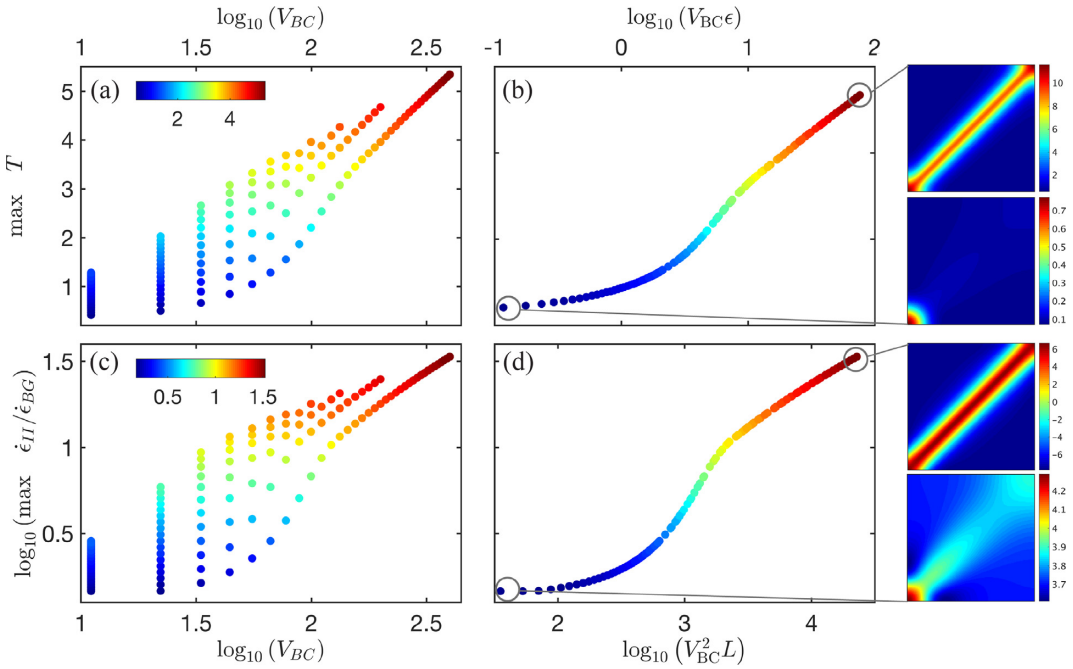

Figure 10. Systematic study of the relative importance of the boundary velocity (V_{BC}) on the maximum temperature and strain rate amplification reached in the shear zone. Panels (a) and (b) depict the maximum temperature and the strain rate amplification factor ($\log_{10} \max(\dot{\epsilon}_{II}/\dot{\epsilon}_{BG})$) achieved in the shear zone as a function of V_{BC}, respectively. The coloured circles represent each individual simulation that was run. The colour bar corresponds to the maximum temperature (in panels a and c) and maximum strain rate amplification (panels b and d) that was reached during the simulations. Panels (c) and (d) show the data collapse for the maximum temperature as a function of $V_{BC}\varepsilon$ and the strain rate amplification factor as a function of $V_{BC}^2 L$. The rightmost subpanels depict temperature and strain rate fields associated with specific parts of the considered parameter space (indicated by the grey lines).

5.2 The development of shear zones in 3-D

To demonstrate the flexibility of the presented PT algorithm, we have extended our GPU code to study 3-D thermomechanical deformation. Two models characterized by different initial thermal conditions were performed. The models were run with a resolution of 158^3 and a relative tolerance of 10^{-5} was achieved at each time step. Both simulations ran for 5000 time steps, and each took about 1 day on a single Nvidia GTX Titan X (Maxwell) GPU card. For both models, we applied zero shear stress on each side and only boundaries normal to the x- and y-axis had a non-zero normal velocity component. Thus, in the first setup, we considered a cylindrical initial thermal perturbation; this configuration is equivalent to the 2-D model discussed above (Section 4.1) and leads to the rapid development of a cylindrical shear zone (Fig. 11a). The second model embedded spherical initial thermal perturbation. For such a configuration, more time is required to propagate the shear zone in 3-D (Fig. 11b). Since more mechanical work is dissipated prior to localization, the shear zone's walls experienced a higher temperature than in the cylindrical case. The maximum temperature in the centre of the shear zones was about 3.5 for both models, which is similar to what was obtained in the 2-D models (Fig. 4).

6 DISCUSSION

6.1 Benefits of the pseudo-transient method and perspectives

The reported results confirm the PT method as performant alternative to a more classical DI solver type to address nonlinear coupled problems in geodynamics. In this study, we proved that both methods are capable of resolving the complex nonlinear physics and converge to an identical solution, even over a large number of time steps. The DI solvers are robust and weakly sensitive to large contrasts in material properties (e.g. viscosity). However, they may require a long and non-trivial development phase. A performant Cholesky factorization of the symmetrical and positive definite finite-difference matrix delivers acceptable time-to-solution for 2-D setups, but is inclined to hit the maximum system wide available RAM memory already at low spatial resolutions for 3-D problems on the considered personal computers: for example, about 60^3 cells. We further report that implicit methods require physical time step values close to CFL if significant nonlinearities are involved.

In contrast, PT solvers result from a simple implementation of the coupled system of equations. The inherent parallelism in PT iterative-based solvers enables a straightforward vectorization, which shows promising implementation on multiple-core accelerators such as GPUs. GPU-based PT solvers outperformed CPU-based DI solvers regarding wall-time, even for 2-D setups. The reported solution of nonlinear thermomechanical problems are identical when computed with either the PT or the DI solver types, and validates the robustness and accuracy of the PT solver implementation. Further, the PT algorithms are succinct codes that enhance readability and make them less error prone. Besides faster times-to-solution, the benefits of PT stencil-based matrix-free solver approaches reside in lower memory footprint, optimal usage of actual hardware and straightforward parallelization, since only neighbours' access is required. Since the investigation of large 3-D setups may require the use of more than a single-GPU accelerator, the GPU-based PT solver can readily be extended to run on a distributed-memory machine, via message-passing interface (MPI). Implementing an MPI point-to-point communication type for subdomain boundary exchange enables the PT solvers to scale on the largest supercomputers and show by construction a close to optimal parallel efficiency (Omlin *et al.* 2017a,b).

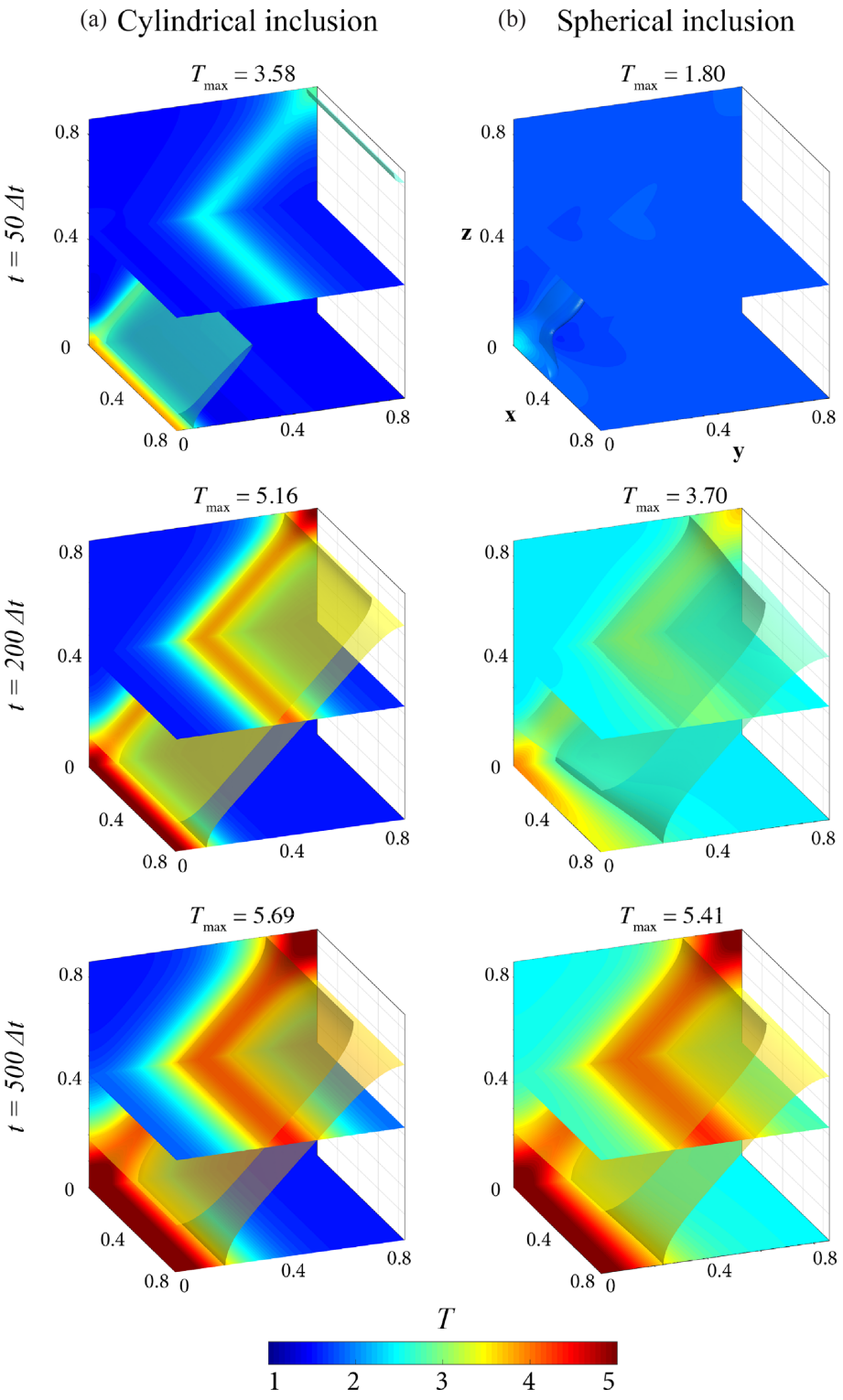

Figure 11. 3-D numerical models of thermomechanical shear localization. Two different configurations were considered: a cylindrical initial thermal perturbation (a) and a spherical initial thermal perturbation. Both perturbations evolve into a single shear zone in response to the mechanical work exerted by the boundaries. Pure shear was applied on boundaries normal to y (inflow) and z (outflow); boundaries normal to x were free to slip. All thermal boundary conditions were insulating.

In this contribution, we focused on the small-strain limit and thus did not treat advection. However, the PT method is not restricted to this specific case and will be extended to convection-type problems in the future. To this end, the PT method will be coupled to either Eulerian (i.e. upwind type) or Eulerian–Lagrangian (i.e. characteristics-based) advection solvers. By analogy with the treatment of nonlinear couplings (see Section 4.6), the PT method will provide a simple framework for implementing either explicit or implicit advection solvers (Furuichi & May 2015). The latter discretization is a method of choice in order to avoid numerical instabilities in convection problems involving a free surface (Furuichi & May 2015).

7 CONCLUSIONS

In the perspective of quantifying and simulating fully coupled thermomechanical processes, such as ductile strain localization owing to shear heating, we have presented two numerical methods based on the finite-difference discretization. The first method is a thermomechanical extension to the DI M2Di solver (Räss *et al.* 2017) that relies on sparse matrix assembly and factorization, and the second method is based on a fully matrix-free PT method. Both methods can model thermomechanically activated shear localization in 2-D and provide consistent results. For 2-D models on a standard desktop computer, the PT method is as accurate and can be as efficient (in terms of wall-time) as the DI approach. We also investigated the impacts of different nonlinear coupling strategies and could show that no matter how much the time step is decreased, solutions obtained with one-way coupling never achieves the accuracy of the two-way coupling.

The significant advantage of the PT method is that it can be extended for high-resolution 3-D numerical simulations without significant modification of the 2-D algorithm and without a drastic increase in memory requirements as the latter scales linearly with the number of grid cells. We show that the PT method is suitable to perform high-resolution 3-D simulations of thermomechanically activated shear localization and other relevant coupled physics in geodynamics (Räss *et al.* 2018). Further, the efficiency of the thermomechanical codes makes it suitable for systematic analysis of the parameters that control the dynamics of shear zone development. Based on 139 2-D simulations, we show that a consistent data collapse of the shear zone temperature and strain rate can be established, which further demonstrate the PT method's robustness.

ACKNOWLEDGEMENTS

The authors acknowledge two anonymous reviewers that provided thorough and constructive reviews of the original draft. The authors are grateful to Daniel Kiss for constructive discussions and Samuel Omlin for his long-term partnership in the high performace code development. We acknowledge the Swiss Geocomputing Centre at the University of Lausanne for support and computing resources. This work was supported by the University of Lausanne.

REFERENCES

Andersen, T.B., Mair, K., Austrheim, H., Podladchikov, Y.Y. & Vrijmoed, J.C., 2008. Stress release in exhumed intermediate and deep earthquakes determined from ultramafic pseudotachylyte, *Geology,* **36**(12), 995–998 .

Braeck, S. & Podladchikov, Y.Y., 2007. Spontaneous thermal runaway as an ultimate failure mechanism of materials, *Phys. Rev. Lett.,* **98,** doi:10.1103/PhysRevLett.98.095504.

Carter, N.L. & Ave'Lallemant, H.G., 1970. High temperature flow of dunite and peridotite, *Bull. geol. Soc. Am.,* **81**(8), 2181–2202.

Choi, E., Tan, E., Lavier, L.L. & Calo, V.M., 2013. DynEarthSol2D: an efficient unstructured finite element method to study long-term tectonic deformation, *J. geophys. Res.,* **118**(5), 2429–2444.

Cundall, P., Coetzee, M., Hart, R. & Varona, P., 1993. *FLAC Users Manual,* Itasca Consulting Group, USA.

Cundall, P.A., 1982. Adaptive density-scaling for time-explicit calculations, in *Proceedings of the 4th International Conference on Numerical Methods in Geomechanics,* pp. 23–26, Edmonton.

Cundall, P.A. & Strack, O.D., 1979. A discrete numerical model for granular assemblies, *Geotechnique,* **29**(1), 47–65.

Duretz, T., Schmalholz, S.M., Podladchikov, Y.Y. & Yuen, D.A., 2014. Physics-controlled thickness of shear zones caused by viscous heating: implications for crustal shear localization, *Geophys. Res. Lett.,* **41**(14), 4904–4911.

Duretz, T., Schmalholz, S.M. & Podladchikov, Y.Y., 2015. Shear heating-induced strain localization across the scales, *Phil. Mag.,* **95**(28–30), 3192–3207.

Eisenstat, S.C., Elman, H.C. & Schultz, M.H., 1983. Variational iterative methods for nonsymmetric systems of linear equations, *SIAM J. Numer. Anal.,* **20**(2), 345–357.

Fleitout, L. & Froidevaux, C., 1980. Thermal and mechanical evolution of shear zones, *J. Struct. Geol.,* **2**(12), 159–164.

Frankel, S.P., 1950. Convergence rates of iterative treatments of partial differential equations, *Math. Comput.,* **4**(30), 65–75.

Furuichi, M. & May, D.A., 2015. Implicit solution of the material transport in stokes flow simulation: toward thermal convection simulation surrounded by free surface, *Comput. Phys. Commun.,* **192,** 1–11.

Gerya, T.V. & Yuen, D.A., 2003. Charactericties-based marker method with conservative finite-difference schemes for modeling geological flows with strongly variable transport properties, *Phys. Earth planet. Inter.,* **140**(4), 293–318.

Gerya, T.V. & Yuen, D.A., 2007. Robust characteristics method for modelling multiphase visco-elasto-plastic thermo-mechanical problems, *Phys. Earth planet. Inter.,* **163**(1-4), 83–105.

Hobbs, B.E. & Ord, A., 1988. Plastic instabilities: implications for the origin of intermediate and deep focus earthquakes, *J. geophys. Res.,* **93**(B9), 10 521–10 540.

Jaquet, Y. & Schmalholz, S.M., 2017. Spontaneous ductile crustal shear zone formation by thermal softening and related stress, temperature and strain rate evolution, *Tectonophysics,* in press.

John, T., Medvedev, S., Rupke, L.H., Andersen, T.B., Podladchikov, Y.Y. & Austrheim, H., 2009. Generation of intermediate-depth earthquakes by self-localizing thermal runaway, *Nat. Geosci.,* **2**(2), 137–140.

Kaus, B.J., Popov, A.A., Baumann, T. & Püsök, A.E., 2016. Forward and inverse modelling of lithospheric deformation on geological timescales, in *Proceedings NIC Symposium 2016,* pp. 299–307 , Jülich, Germany.

Kaus, B.J.P. & Podladchikov, Y.Y., 2006. Initiation of localized shear zones in viscoelastoplastic rocks, *J. geophys. Res.,* **111**(B4), doi:10.1029/2005JB003652.

McKenzie, D.P., Roberts, J.M. & Weiss, N.O., 1974. Convection in the Earth's mantle: towards a numerical simulation, *J. Fluid Mech.,* **62**(3), 465–538.

Moore, J.D. & Parsons, B., 2015. Scaling of viscous shear zones with depth-dependent viscosity and power-law stress-strain-rate dependence, *Geophys. J. Int.,* **202**(1), 242–260.

Ogawa, M., 1987. Shear instability in a viscoelastic material as the cause of deep focus earthquakes, *J. geophys. Res.,* **92**(B13), 13 801–13 810.

Ohuchi, T., Lei, X., Ohfuji, H., Higo, Y., Tange, Y., Sakai, T., Fujino, K. & Irifune, T., 2017. Intermediate-depth earthquakes linked to localized heating in dunite and harzburgite, *Nat. Geosci.,* **10,** 771–776.

Omlin, S., 2017. Development of massively parallel near peak performance solvers for three-dimensional geodynamic modelling, *PhD thesis,* University of Lausanne.

Omlin, S., Malvoisin, B. & Podladchikov, Y.Y., 2017a. Pore fluid extraction by reactive solitary waves in 3-D, *Geophys. Res. Lett.,* **44,** 9267–9275.

Omlin, S., Räss, L. & Podladchikov, Y.Y., 2017b. Simulation of three-dimensional viscoelastic deformation coupled to porous fluid flow, *Tectonophysics,* in press.

Otter, J. R.H., Cassell, A.C. & Hobbs, R.E., 1966. Dynamic relaxation, *Proc. Inst. Civ. Eng.,* **35**(4), 633–656.

Parsons, B. & McKenzie, D., 1978. Mantle convection and the thermal structure of the plates, *J. geophys. Res.,* **83**(B9), 4485–4496.

Pekeris, C.L., 1935. Thermal convection in the interior of the Earth, *Geophys. J. Int.,* **3,** 343–367.

Peters, M., Veveakis, M., Poulet, T., Karrech, A., Herwegh, M. & Regenauer-Lieb, K., 2015. Boudinage as a material instability of elasto-visco-plastic rocks, *J. Struct. Geol.,* **78,** 86–102.

Poliakov, A.N.B., Cundall, P.A., Podladchikov, Y.Y. & Lyakhovsky, V.A., 1993. An explicit inertial method for the simulation of viscoelastic flow: an evaluation of elastic effects on diapiric flow in two- and three-layers models, in *Flow and Creep in the Solar System: Observations, Modeling and Theory,* pp. 175–195, eds Stone, D.B. & Runcorn, S.K., Springer Netherlands.

Popov, A. & Sobolev, S., 2008. Slim3D: a tool for three-dimensional thermomechanical modeling of lithospheric deformation with elasto-visco-plastic rheology, *Phys. Earth planet. Inter.,* **171**(1–4), 55–75.

Prieto, G.A., Florez, M., Barrett, S.A., Beroza, G.C., Pedraza, P., Blanco, J.F. & Poveda, E., 2013. Seismic evidence for thermal runaway during intermediate-depth earthquake rupture, *Geophys. Res. Lett.,* **40**(23), 6064–6068.

Räss, L., Duretz, T., Podladchikov, Y.Y. & Schmalholz, S.M., 2017. M2Di: concise and efficient MATLAB 2-D Stokes solvers using the Finite Difference Method, *Geochem. Geophys. Geosyst.,* **18**(2), 755–768.

Räss, L., Simon, N.S. & Podladchikov, Y.Y., 2018. Spontaneous formation of fluid escape pipes from subsurface reservoirs, *Sci. Rep.,* **8**(1), doi:10.1038/s41598-018-29485-5.

Regenauer-Lieb, K., Yuen, D.A. & Branlund, J., 2001. The initiation of subduction: criticality by addition of water?, *Science,* **294**(5542), 578–580.

Rice, A. & Fairbridge, R., 1975. Thermal runaway in the mantle and neotectonics, *Tectonophysics,* **29**(1), 59–72.

Schmalholz, S.M. & Duretz, T., 2015. Shear zone and nappe formation by thermal softening, related stress and temperature evolution, and application to the Alps, *J. Metamorphic Geol.,* **33**(8), 887–908.

Takeuchi, C.S. & Fialko, Y., 2012. Dynamic models of interseismic deformation and stress transfer from plate motion to continental transform faults, *J. geophys. Res.,* **117**(B5), doi:10.1029/2011JB009056.

Thielmann, M. & Kaus, B.J., 2012. Shear heating induced lithospheric-scale localization: does it result in subduction?, *Earth planet. Sci. Lett.,* **359-360,** 1–13.

Thielmann, M., Rozel, A., Kaus, B. & Ricard, Y., 2015. Intermediate-depth earthquake generation and shear zone formation caused by grain size reduction and shear heating, *Geology,* **43**(9), 791–794.

Wilson, C.R., Spiegelman, M. & van Keken, P.E., 2017. TerraFERMA: the Transparent Finite Element Rapid Model Assembler for multiphysics problems in Earth sciences, *Geochem. Geophys. Geosyst.,* **18,** 769–810.

Yang, X.I. & Mittal, R., 2014. Acceleration of the Jacobi iterative method by factors exceeding 100 using scheduled relaxation, *J. Comput. Phys.,* **274,** 695–708.

Yuen, D.A. & Schubert, G., 1979. On the stability of frictionally heated shear flows in the asthenosphere, *Geophys. J. Int.,* **57**(1), 189–207.

APPENDIX A: NON-DIMENSIONALIZATION OF THE THERMOMECHANICAL EQUATIONS

Dimensionless thermomechanical equations (eq. 4) can be obtained by introducing the characteristic scales (eq. 3) in the dimensional thermomechanical eq. (1). The dimensionless equations are obtained by substituting all dimensional variables by the product of their characteristic values and dimensionless value. For example, the temperature can be expressed as $T = \bar{T}\tilde{T}$, there the tilde sign stands for the dimensionless temperature.

Here the temperature evolution equation is expressed as

$$0 = \tau_{ij}\dot{\epsilon}_{ij} + k\frac{\partial^2 T}{\partial x_i^2} - \rho C_\mathrm{p}\frac{\partial T}{\partial t} = \frac{\bar{\tau}}{\bar{t}}\tilde{\tau}_{ij}\tilde{\dot{\epsilon}}_{ij} + k\frac{\bar{T}}{\bar{L}^2}\frac{\partial^2 \tilde{T}}{\partial \tilde{x}_i^2} - \rho C_\mathrm{p}\frac{\bar{T}}{\bar{t}}\frac{\partial \tilde{T}}{\partial \tilde{t}}. \quad \text{(A1)}$$

By dividing the above expression by $\bar{\tau}/\bar{t}$ yields

$$0 = \tilde{\tau}_{ij}\tilde{\dot{\epsilon}}_{ij} + k\frac{\bar{t}\bar{T}}{\bar{\tau}\bar{L}^2}\frac{\partial^2 \tilde{T}}{\partial \tilde{x}_i^2} - \rho C_\mathrm{p}\frac{\bar{T}}{\bar{\tau}}\frac{\partial \tilde{T}}{\partial \tilde{t}}. \quad \text{(A2)}$$

After introducing the characteristic length, $\bar{L} = \sqrt{\frac{k}{\rho C_\mathrm{p}}\bar{t}}$ and stress, $\bar{\tau} = \rho C_\mathrm{p}\bar{T}$, one finally gets

$$0 = \tilde{\tau}_{ij}\tilde{\dot{\epsilon}}_{ij} + \frac{\partial^2 \tilde{T}}{\partial \tilde{x}_i^2} - \frac{\partial \tilde{T}}{\partial \tilde{t}}. \quad \text{(A3)}$$

The dimensional constitutive relationship is spelled as

$$\dot{\epsilon}_{ij} = \frac{1}{2}A\tau_\mathrm{II}^{n-1}\exp\left(-\frac{Q}{R\,(T_0+T)}\right)\tau_{ij}. \quad \text{(A4)}$$

One first introduces the characteristic time and stress to express the dimensionless strain rate:

$$\frac{\tilde{\dot{\epsilon}}_{ij}}{\bar{t}} = 2^{-1}A\bar{\tau}^{n-1}\tilde{\tau}_\mathrm{II}^{n-1}\exp\left(-\frac{Q}{R\,(T_0+T)}\right)\bar{\tau}_{ij}\tilde{\tau}_{ij}, \quad \text{(A5)}$$

which can be recasted as

$$\tilde{\dot{\epsilon}}_{ij} = \bar{t}\,2^{-1}A\bar{\tau}^{n}\tilde{\tau}_\mathrm{II}^{n-1}\exp\left(-\frac{Q}{R\,(T_0+T)}\right)\tilde{\tau}_{ij}. \quad \text{(A6)}$$

After substitution of the characteristic time, $\bar{t} = 2^{1-n}A^{-1}\bar{\tau}^{-n}\exp\left(\frac{Q}{RT_0}\right)$, the expression simplifies to

$$\tilde{\dot{\epsilon}}_{ij} = 2^{-n}\tilde{\tau}_\mathrm{II}^{n-1}\exp\left(-\frac{Q}{R\,(T_0+T)}+\frac{Q}{RT_0}\right)\tilde{\tau}_{ij}. \quad \text{(A7)}$$

Expressing the Arrhenius term with a common denominator and introducing the dimensionless temperature yields:

$$\tilde{\dot{\epsilon}}_{ij} = 2^{-n}\tilde{\tau}_\mathrm{II}^{n-1}\exp\left(\frac{QR\bar{T}\tilde{T}}{R^2T_0^2+R^2T_0\bar{T}\tilde{T}}\right)\tilde{\tau}_{ij}. \quad \text{(A8)}$$

By substituting , $\bar{T} = \frac{nRT_0^2}{Q}$ at the numerator and dividing both the numerator and denominator by $R^2T_0^2$ leads to

$$\tilde{\dot{\epsilon}}_{ij} = 2^{-n}\tilde{\tau}_\mathrm{II}^{n-1}\exp\left(\frac{n\tilde{T}}{1+\frac{\bar{T}}{T_0}\tilde{T}}\right)\tilde{\tau}_{ij}. \quad \text{(A9)}$$

Finally one may introduce a dimensionless reference temperature, $\tilde{T}_0 = T_0/\bar{T}$, which simplifies the expression in the following way:

$$\tilde{\dot{\epsilon}}_{ij} = 2^{-n}\tilde{\tau}_\mathrm{II}^{n-1}\exp\left(\frac{n\tilde{T}}{1+\frac{\tilde{T}}{\tilde{T}_0}}\right)\tilde{\tau}_{ij}. \quad \text{(A10)}$$

The dimensionalization of the momentum equations (in the absence of body forces) and of the continuity equation is straightforward, hence they will not be detailed here. Please note that in the main body of the text, we omit the ˜ superscript to enhance the readability.

APPENDIX B: SOLVING THE NONLINEAR THERMOMECHANICAL SYSTEM WITH THE DIRECT-ITERATIVE SCHEME

An implicit (or semi-implicit) discretization (e.g. backward-Euler, Crank–Nicolson) results in a linear system of equation of the form:

$$\mathbf{K}_{\mathrm{TM}}\,\mathbf{x} = \mathbf{b}, \tag{B1}$$

The systems of equation couples the different solution fields $\mathbf{v}$, $\mathbf{p}$, $\mathbf{T}$ in the following way:

$$\underbrace{\begin{bmatrix} \mathbf{K}_{\mathrm{vv}} & \mathbf{K}_{\mathrm{vp}} & \mathbf{0} \\ \mathbf{K}_{\mathrm{pv}} & \mathbf{0} & \mathbf{0} \\ \mathbf{K}_{\mathrm{Tv}} & \mathbf{0} & \mathbf{K}_{\mathrm{TT}} \end{bmatrix}}_{\mathbf{K}_{\mathrm{TM}}} \underbrace{\begin{bmatrix} \mathbf{v} \\ \mathbf{p} \\ \mathbf{T} \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} \mathbf{b}_{\mathrm{v}} \\ \mathbf{b}_{\mathrm{p}} \\ \mathbf{b}_{\mathrm{T}} \end{bmatrix}}_{\mathbf{b}}, \tag{B2}$$

where the matrices $\mathbf{K}_{\mathrm{TM}}$ contains coefficients resulting from the discretization (here, centred finite differences). The blocks $\mathbf{K}_{\mathbf{vv}}$ and $\mathbf{K}_{\mathbf{Tv}}$ represent the deviatoric stress gradient and shear heating operators, respectively. The $\mathbf{K}_{\mathbf{vp}}$ block represents the discrete gradient operator, which equals minus the transpose of the divergence operator, that is, $\mathbf{K}_{\mathbf{pv}} = -\mathbf{K}_{\mathrm{vp}}{}^{T}$. The $\mathbf{K}_{\mathbf{TT}}$ is the Laplace operator, which arises from the diffusion term in the temperature evolution equation. The $\mathbf{b}$ contains contributions resulting from the time discretization of the transient terms (here in $\mathbf{b}_{\mathrm{T}}$) as well as contributions from the boundary conditions (e.g. Dirchlet and Newman boundary values).

In the case of a linear problem, the solutions $\mathbf{x}$ can readily be obtained by applying the inverse of the matrix $\mathbf{K}_{\mathrm{TM}}$ to the right-hand side vector $\mathbf{b}$:

$$\mathbf{x} = \mathbf{K}_{\mathrm{TM}}{}^{-1}\,\mathbf{b}. \tag{B3}$$

In the present case, the nonlinear dependence of the viscosity on the temperature and the velocity (i.e. via the strain rate dependence) causes the global system of equation to be nonlinear. The nonlinearity results in an imbalance of the global equations:

$$\mathbf{b} - \mathbf{K}_{\mathrm{TM}}\,\mathbf{x} = \mathbf{f} \neq 0, \tag{B4}$$

where $\mathbf{f}$ is the nonlinear residual vector. In order to minimize the imbalance, quantified by the magnitude $\mathbf{f}$, it is convenient to solve this problem iteratively. To this end, the solution vector is updated within a cycle of k nonlinear iterations:

$$\mathbf{x}^{k+1} = \mathbf{x}^{k} + \alpha\delta\mathbf{x}^{k+1}, \tag{B5}$$

where $\delta\mathbf{x}^{k}$ is the nonlinear correction to the solution. The scalar parameter α is determined during a line search procedure in order to minimize the magnitude of the residuals:

$$\min \left|\left|\mathbf{f}\left(\mathbf{x}^{k} + \alpha\ \delta\mathbf{x}^{k+1}\right)\right|\right|_{\mathrm{L2}}. \tag{B6}$$

In the context of Picard iterations, these corrections are obtained by evaluating the matrix $\mathbf{K}_{\mathrm{TM}}$ at each iteration, using the values of viscosities evaluated with the current temperature and velocity fields, and by computing the correction as

$$\delta\mathbf{x}^{k+1} = \left(\mathbf{K}_{\mathrm{TM}}{}^{k}\right)^{-1}\,\mathbf{f}^{k}, \tag{B7}$$

where k is the nonlinear iteration index. The vector $\mathbf{f}$ can be obtained with the following operation:

$$\underbrace{\begin{bmatrix} \mathbf{f}_{\mathrm{v}} \\ \mathbf{f}_{\mathrm{p}} \\ \mathbf{f}_{\mathrm{T}} \end{bmatrix}}_{\mathbf{f}^{k}} = \underbrace{\begin{bmatrix} \mathbf{b}_{\mathrm{v}} \\ \mathbf{b}_{\mathrm{p}} \\ \mathbf{b}_{\mathrm{T}} \end{bmatrix}}_{\mathbf{b}^{k}} - \underbrace{\begin{bmatrix} \mathbf{K}_{\mathbf{vv}}(\mathbf{v},\mathbf{T}) & \mathbf{K}_{\mathrm{vp}} & \mathbf{0} \\ \mathbf{K}_{\mathrm{pv}} & \mathbf{0} & \mathbf{0} \\ \mathbf{K}_{\mathbf{Tv}}(\mathbf{v},\mathbf{T}) & \mathbf{0} & \mathbf{K}_{\mathrm{TT}} \end{bmatrix}}_{\mathbf{K}_{\mathrm{TM}}{}^{k}} \underbrace{\begin{bmatrix} \mathbf{v} \\ \mathbf{p} \\ \mathbf{T} \end{bmatrix}}_{\mathbf{x}^{k}}, \tag{B8}$$

Here, we emphasize the nonlinearity held in the blocks $\mathbf{K}_{\mathbf{vv}}$ and $\mathbf{K}_{\mathbf{Tv}}$ due to the nonlinear dependence of viscosity on the velocity and temperature. Nonlinear iterations need to be performed until the magnitude of the nonlinear residual has decreased below a given level of tolerance, for example, $\|\mathbf{f}^{k}\|_{2} < \mathrm{tol}$. The Picard iterations only deliver a linear rate of convergence and therefore, a high number of iterations is generally required before reaching the desired nonlinear accuracy (i.e. several tens of iterations).

To overcome this severe restriction, we have used a Newton linearization. To this end, the iteration matrix in eq. (B11) is substituted by the Jacobian matrix $\mathbf{J}_{\mathrm{TM}}$. The Jacobian matrix takes the form of

$$\mathbf{J}_{\mathrm{TM}} = \frac{\partial \mathbf{f}_{i}}{\partial \mathbf{x}_{j}} \tag{B9}$$

and thus contains information about the gradient of the residual with regard to the solution. In practice, for the considered case, the Jacobian matrix can be written as

$$\mathbf{J}_{\mathrm{TM}} = \begin{bmatrix} \mathbf{J}_{\mathrm{vv}} & \mathbf{K}_{\mathrm{vp}} & \mathbf{J}_{\mathrm{vT}} \\ \mathbf{K}_{\mathrm{pv}} & \mathbf{0} & \mathbf{0} \\ \mathbf{J}_{\mathrm{Tv}} & \mathbf{0} & \mathbf{J}_{\mathrm{TT}} \end{bmatrix}, \tag{B10}$$

where the blocks $\mathbf{J}_{\mathrm{vv}}$ and $\mathbf{J}_{\mathrm{Tv}}$ differ from $\mathbf{K}_{\mathbf{vv}}$ and $\mathbf{K}_{\mathbf{Tv}}$ since they contain additional contributions from the gradients of viscosity with regard to the velocity. The new block $\mathbf{J}_{\mathrm{vT}}$ as well as the block $\mathbf{J}_{\mathbf{TT}}$ contain information from the gradients of viscosity with regard to the temperature. The structure of the matrix operator $\mathbf{J}_{\mathrm{TM}}$ shares similarities with that employed by Wilson *et al.* (2017), who applied similar linearization to study thermomechanical convection. The Newton correction is hence obtained with the following operation:

$$\delta\mathbf{x}^{k+1} = \left(\mathbf{J}_{\mathrm{TM}}{}^{k}\right)^{-1}\,\mathbf{f}^{k}, \tag{B11}$$

which allows to reach the desired nonlinear accuracy with a quadratic rate of convergence (i.e. less than 10 iterations).

APPENDIX C: THE DIRECT-ITERATIVE SOLVER

We seek a solution of the following linear system:

$$\underbrace{\begin{bmatrix} \mathbf{J}_{\mathrm{vv}} & \mathbf{K}_{\mathrm{vp}} & \mathbf{J}_{\mathrm{vT}} \\ \mathbf{K}_{\mathrm{pv}} & \mathbf{0} & \mathbf{0} \\ \mathbf{J}_{\mathrm{Tv}} & \mathbf{0} & \mathbf{J}_{TT} \end{bmatrix}}_{\mathbf{J}_{\mathrm{TM}}} \underbrace{\begin{bmatrix} \delta\mathbf{v} \\ \delta\mathbf{p} \\ \delta\mathbf{T} \end{bmatrix}}_{\delta\mathbf{x}} = -\underbrace{\begin{bmatrix} \mathbf{f}_{\mathrm{v}} \\ \mathbf{f}_{\mathrm{p}} \\ \mathbf{f}_{\mathrm{T}} \end{bmatrix}}_{\mathbf{f}}. \tag{C1}$$

To facilitate and enhance the linear solve procedure, we introduce the pre-conditionner:

$$\mathbf{J}_{\mathrm{TM}}^{\mathrm{pc}} = \begin{bmatrix} \mathbf{J}_{\mathrm{vv}} & \mathbf{K}_{\mathrm{vp}} & \mathbf{J}_{\mathrm{vT}} \\ \mathbf{K}_{\mathrm{pv}} & \mathbf{J}_{\mathrm{pp}} & \mathbf{0} \\ \mathbf{J}_{\mathrm{Tv}} & \mathbf{0} & \mathbf{J}_{\mathrm{TT}} \end{bmatrix}, \tag{C2}$$

where the block matrix $\mathbf{J}_{\mathrm{pp}} = \gamma^{-1}\mathbf{I}$ corresponds to a weakly compressible contribution.

The linear residuals are defined as

$$\begin{cases} \mathbf{r}_{\mathrm{v}} = \mathbf{f}_{\mathrm{v}} - \mathbf{J}_{\mathrm{vv}}\delta\mathbf{v} - \mathbf{K}_{\mathrm{vp}}\delta\mathbf{p} - \mathbf{J}_{\mathrm{vT}}\delta\mathbf{T} \\ \mathbf{r}_{\mathrm{p}} = \mathbf{f}_{\mathrm{p}} - \mathbf{K}_{\mathrm{pv}}\delta\mathbf{v} \\ \mathbf{r}_{\mathrm{T}} = \mathbf{f}_{\mathrm{T}} - \mathbf{J}_{\mathrm{Tv}}\delta\mathbf{v} - \mathbf{J}_{\mathrm{TT}}\delta\mathbf{T} \end{cases}. \tag{C3}$$

The solutions are found iteratively:

$$\begin{cases} \delta\mathbf{v}^{i+1} = \delta\mathbf{v}^{i} + \delta\delta\mathbf{v} \\ \delta\mathbf{p}^{i+1} = \delta\mathbf{p}^{i} + \delta\delta\mathbf{p} \\ \delta\mathbf{T}^{i+1} = \delta\mathbf{T}^{i} + \delta\delta\mathbf{T}, \end{cases} \tag{C4}$$

where $\delta\delta\mathbf{v}$, $\delta\delta\mathbf{p}$ and $\delta\delta\mathbf{T}$ are iterative corrections and i is the iteration count.

The iterative velocity correction is obtained by solving:

$$\delta\delta\mathbf{v} = \bar{\mathbf{J}}_{\mathrm{vv}}{}^{-1}\,\bar{\mathbf{r}}_{\mathrm{v}}, \tag{C5}$$

where $\bar{\mathbf{J}}_{\mathrm{vv}} = \mathbf{J}_{\mathrm{vv}} - \mathbf{K}_{\mathrm{vp}}\left(\mathbf{J}_{\mathrm{pp}}{}^{-1}\,\mathbf{K}_{\mathrm{pv}}\right)$ and $\bar{\mathbf{r}}_{\mathrm{v}} = \mathbf{r}_{\mathrm{v}} - \mathbf{K}_{\mathrm{vp}}\left(\mathbf{J}_{\mathrm{pp}}{}^{-1}\,\mathbf{r}_{\mathrm{p}}\right) - \mathbf{K}_{\mathbf{vT}}\delta\delta\mathbf{T}$. Applying the inverse of $\mathbf{J}_{\mathrm{pp}}$ is a trivial operation, since $\mathbf{J}_{\mathbf{pp}}$ is a diagonal block matrix. However, applying the inverse of $\bar{\mathbf{J}}_{\mathrm{vv}}$ is a cumbersome task, since $\bar{\mathbf{J}}_{\mathrm{vv}}$ may not be a symmetrical matrix. Here, we use an iterative Krylov subspace solver (generalized conjugate residuals, Eisenstat *et al.* 1983) where the Cholesky factors of the symmetrical operators resulting from a Picard linearization are used for preconditioning. This approach is described in detail by Räss *et al.* (2017).

Subsequently, the iterative pressure correction is obtained in a trivial way by evaluating

$$\delta\delta\mathbf{p} = \mathbf{J}_{\mathrm{pp}}{}^{-1}\left(\mathbf{r}_{\mathrm{p}} - \mathbf{K}_{\mathrm{pv}}\delta\delta\mathbf{v}\right). \tag{C6}$$

Finally, the iterative temperature correction is calculated as follows:

$$\delta\delta\mathbf{T} = \mathbf{J}_{\mathrm{TT}}{}^{-1}\left(\mathbf{r}_{\mathrm{T}} - \mathbf{J}_{\mathrm{Tv}}\delta\delta\mathbf{v}\right). \tag{C7}$$

Here, $\mathbf{J}_{\mathrm{TT}}$ is a symmetric positive definite matrix and its inverse can be efficiently applied using pre-computed Cholesky factors. The iteration is repeated until the L_2 norm of all linear residual vectors has decreased below a given threshold value.