https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

TERRA-NG v1.0: Extreme-Scale, GPU-accelerated Mantle Convection

Fabian Böhm[1], Nils Kohl[3], Ponsuganth Ilangovan[3], Gabriel Robl[3], Fatemeh Rezaei[3], Marcus Mohr[3], Bernhard S. A. Schuberth[3], Harald Köstler[1,2], Hans-Peter Bunge[3], and Ulrich Rüde[2,4]

[1]Erlangen National High Performance Computing Center (NHR@FAU), Erlangen, Germany
[2]Chair for System Simulation (CS10), Friedrich–Alexander–Universität Erlangen–Nürnberg, Erlangen, Germany
[3]Department of Earth and Environmental Sciences, LMU Munich, Germany
[4]CERFACS, Toulouse, France

Correspondence: Fabian Böhm (fabian.boehm@fau.de)

Abstract. We present TERRA-NG, a portable, GPU-accelerated, matrix-free mantle-convection code. A single Kokkos C++ implementation runs at scale on NVIDIA, AMD, and Intel GPU supercomputers. TERRA-NG has a deliberately narrow design: built on a radially extruded mesh of spherical wedges, tailored to the spherical shell geometry, which enables domain-specific optimizations like single quadrature-point integral-evaluations, radial coordinate storage compression and radial shared-memory tiling. The corresponding low-order W_1-iso-W_2/W_1 wedge-based Stokes–energy discretisation is verified against the Zhong et al. (2008) spherical-shell convection benchmark suite. We showcase TERRA-NG through strong- and weak-scaling on the JUWELS Booster (NVIDIA A100), MareNostrum 5 (NVIDIA H100), LUMI-G (AMD MI250X), Hunter (AMD MI300A APU), and SuperMUC-NG Phase 2 (Intel PVC) supercomputers. Coupled mantle convection simulations at ~ 11 km and ~ 5.6 km radial spacing (~ 2.8 B and ~ 22 B DoFs) can be run routinely on standard node partitions of all considered systems. Global ~ 1 km-per-gridpoint mantle convection (~ 1.4 T DoFs) is feasible on an extreme-scale allocation, and a sub-km hero-run at ~ 0.7 km grid spacing scaling up to $\sim 11,000$ GPUs of LUMI-G (~ 11 T DoFs) shows the potential of the code on future, larger machines.

Introduction

The Earth's mantle is a roughly 3000 km thick layer of rock between the Earth's crust and the core and the largest part of the planet in terms of volume. Although stronger than steel on the time scale of hours, it deforms and flows over geologic time. This creeping motion is known as mantle convection. It is the driver of all large-scale tectonic activity on our planet, including plate tectonics and mountain building, and understanding it is essential to understanding the evolution of Earth and other terrestrial planets (Davies, 1999; Schubert et al., 2001).

Due to the high viscosity of mantle rock, on the order of 10^{21} Pa s, inertial forces are negligible relative to viscous forces, and the Reynolds number of the flow is effectively zero. In this creeping-flow (Stokes) regime the instantaneous velocity field is entirely determined by the buoyancy forces and boundary conditions, and Rayleigh number and geometry control the pattern, wavelength, and amplitude of convective motion. This makes it important to model mantle convection in its 3-D spherical-

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

shell geometry. At the same time, the mantle convects at very high Rayleigh number, resulting in thin thermal boundary layers, narrow upwelling plumes and downwelling slabs, and a broad range of interacting scales (Schuberth et al., 2009; Gassmöller et al., 2020; Ilangovan et al., 2026). High numerical resolution in 3-D spherical geometry is therefore essential for capturing the dynamics of the convecting mantle.

Early 3-D spherical mantle convection codes were developed in the mid-1980s, with the pioneering work of Baumgardner (1985) and Glatzmaier (1988). The former introduced an icosahedral discretization of the spherical shell (Baumgardner and Frederickson, 1985) that provides a nearly uniform resolution over the sphere. The resulting TERRA code was among the first geodynamics codes to be parallelized for distributed-memory compute clusters (Bunge and Baumgardner, 1995), producing key results on the role of depth-dependent viscosity and the resulting large-scale convective planform of the mantle (Bunge et al., 1996, 1997).

A diverse ecosystem of mantle-convection codes has emerged over the past four decades. Alongside TERRA, the community established CitcomS on a 12-cap spherical mesh (Zhong et al., 2000, 2008), StagYY on the yin-yang grid (Tackley, 2008), Underworld for particle-in-cell rheology (Moresi et al., 2007), Fluidity for adaptive unstructured discretizations (Davies et al., 2011), pTatin3D for matrix-free lithospheric dynamics (May et al., 2014, 2015), the octree-adaptive Rhea code (Burstedde et al., 2013), ASPECT (Kronbichler et al., 2012; Heister et al., 2017), which leverages the deal.II finite-element framework, and G-ADOPT (Davies et al., 2022), built on the automated finite-element system Firedrake (Rathgeber et al., 2016). Stokes solves for mantle convection on pre-exascale CPU systems reached six hundred billion degrees of freedom on 1.6 million cores in the 2015 Gordon Bell prize-winning work of Rudi et al. (2015). TERRA itself was rebuilt on the HyTeG hierarchical hybrid grid framework as part of the TerraNeo project (Kohl et al., 2019; Bauer et al., 2019, 2020; Ilangovan et al., 2026). GPU-targeted work for geodynamic Stokes problems has so far been more limited in scope (Zheng et al., 2014; Kronbichler and Ljungkvist, 2019; Clevenger and Heister, 2021; Lee et al., 2026).

Despite such advances, the attainable resolutions remain below what is needed to resolve the full range of dynamically active scales and the interaction of convection with realistic, strongly temperature- and stress-dependent rheologies. For instance, the Earth's surface area is $4\pi R_o^2 \approx 5.1 \times 10^8\,\mathrm{km}^2$ ($R_o = 6371\,\mathrm{km}$). An assumed icosahedral surface grid at refinement level ℓ consisting of 10 diamonds holds roughly $N_{\mathrm{lat}} = 10 \times 4^\ell$ nodes, yielding a lateral spacing of $h \approx \sqrt{4\pi R_o^2/N_{\mathrm{lat}}}$. At level 13, $N_{\mathrm{lat}} \approx 6.7 \times 10^8$ and $h \approx 0.87\,\mathrm{km}$. Matching this resolution radially across the 2891 km deep shell requires $N_{\mathrm{rad}} \approx 3300$ layers, giving $N = N_{\mathrm{rad}} \cdot N_{\mathrm{lat}} \approx 2.2 \times 10^{12}$ nodes and $\sim 1.1 \times 10^{13}$ degrees of freedom (DoFs; assuming 5 unknowns per node: 3 velocity + 1 pressure + 1 temperature) for the momentum and energy balance. A single solution vector at this resolution with double accuracy occupies $\sim 9 \times 10^{13}$ bytes ($\approx 88\,\mathrm{TB}$). Storing a sparse matrix with a simple Laplacian stencil of $\sim$15 entries per row would require $\sim 2.6 \times 10^{15}$ bytes ($\approx 2.6\,\mathrm{PB}$, 64-bit values and column indices), filling and exceeding the total GPU memory of contemporary top tier machines (Table 3: LUMI-G aggregates $\sim 1.5\,\mathrm{PB}$ of HBM). Matrix-free methods (May et al., 2015; Kronbichler and Kormann, 2012), which compute the effect of an operator application without assembling the associated global matrix, are therefore essential. Furthermore, given the immense computational demands of the application, scalability and node-level performance are major concerns in mantle convection simulations. They are not an incremental upgrade for these simulations but the precondition that makes them feasible.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

Simulations of this scale can only be conducted on supercomputers, where the hardware landscape has shifted decisively toward GPU-accelerated architectures, with all three major vendors competing: NVIDIA powers systems such as JUWELS Booster and JUPITER (Germany), MareNostrum 5 (Spain), Summit (USA), and ABCI (Japan); AMD equips Frontier and El Capitan (USA), LUMI-G (Finland), and Setonix (Australia); and Intel provides the GPUs for SuperMUC-NG Phase 2 (Germany) and Aurora (USA). The resulting heterogeneous landscape means that application codes must be portable across GPU backends to exploit the available resources.

Contribution

In this paper, we present a new finite element code for 3-D spherical-shell mantle convection designed from the outset for GPU architectures. By exploiting the massive parallelism of modern GPUs, our code achieves numerical resolutions substantially exceeding those of previous spherical mantle convection models, opening the possibility of directly resolving convective structures at a level of detail previously inaccessible to global-scale simulations. TERRA-NG draws its numerical design from TERRA (Baumgardner, 1985; Baumgardner and Frederickson, 1985; Bunge and Baumgardner, 1995; Davies et al., 2013): a single-application code for mantle convection on a radially extruded icosahedral mesh of spherical wedges, a geometry to which we tailor kernel-level optimisations that would not carry over to a general FE setting. Through Kokkos, a single source compiles unchanged across CUDA, HIP, SYCL, and multicore CPU backends and, paired with GPU-aware MPI, scales on NVIDIA, AMD, and Intel GPU supercomputers without per-vendor specialisation.

This paper presents the software architecture, the spherical-wedge Finite Element discretisation and its verification against spherical mantle convection references, as well as a feasibility study of the coupled Stokes–energy simulation across five GPU supercomputers (Table 3). In an extreme-scale access, TERRA-NG reaches the unprecedented resolution of $\sim$1 km per grid point in a global mantle-convection simulation for the first time, at timestep costs that would, given a sufficient allocation, permit such runs routinely. A companion paper (Böhm et al., 2026c) presents the cross-vendor node-level performance analysis of the compute kernels, which enabled reaching this resolution goal.

Throughout the paper, mesh resolution is parameterised by an integer level MT on the radially extruded icosahedral mesh, whose construction and data structure are described in detail in Section 2: each diamond edge of the icosahedral shell carries MT lateral cells, and the shell is extruded through MT/2 radial layers, so e.g. MT128 corresponds to 128 lateral cells per edge and 64 radial layers. Table 1 lists the grid sizes, the implied Earth-mantle spacings, and the corresponding DoF counts for the W_1-iso-W_2/W_1 Stokes discretisation. The lateral spacing is h_{lat}, tabulated at $R_{\mathrm{surf}} = 6371$ km and $R_{\mathrm{CMB}} = 3480$ km, and the radial spacing is h_{rad}. The sweep spans a range from MT32 ($h_{\mathrm{rad}} \approx 181$ km, ~ 0.7 M DoFs, well coarser than any geodynamically meaningful flow model) to MT4096 ($h_{\mathrm{rad}} \approx 1.4$ km, ~ 1.4 T DoFs, at the resolution regime targeted in the introduction) up to sub 1km-scale with MT8192.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Per-macro-edge grid sizes, implied Earth-mantle resolution (assuming shell thickness $D = 2891$ km), and total DoF count for the W_1-iso-W_2/W_1 Stokes discretisation across the MT sweep. MT4096 is highlighted as the target of this work: it is the coarsest level reaching the kilometre-scale grid spacing that global mantle convection requires.

MT	lat cells	rad layers	$h_{\mathrm{lat}}^{\mathrm{surf}}$ [km]	$h_{\mathrm{lat}}^{\mathrm{CMB}}$ [km]	h_{rad} [km]	DoF total
MT32	32	16	220.4	120.4	180.7	0.72 M
MT64	64	32	110.2	60.2	90.3	5.58 M
MT128	128	64	55.1	30.1	45.2	43.9 M
MT256	256	128	27.6	15.1	22.6	349 M
MT512	512	256	13.8	7.5	11.3	2.78 B
MT1024	1024	512	6.9	3.8	5.6	22.2 B
MT2048	2048	1024	3.4	1.9	2.8	177 B
MT4096	**4096**	**2048**	**1.7**	**0.9**	**1.4**	**1.42 T**
MT8192	8192	4096	0.85	0.46	0.71	11.4 T

1 Model

Mantle convection is governed by the conservation of momentum (Stokes equations), conservation of mass (anelastic constraint), and an energy balance (advection–diffusion of temperature with adiabatic, shear and internal heating), coupled through thermal buoyancy and temperature-dependent rheology (Schubert et al., 2001). Mantle-convection models differ in how they treat compressibility, ranging from the Boussinesq approximation over the anelastic liquid approximation (Gough, 1969; Jarvis and McKenzie, 1980) to the fully compressible case, see Gassmöller et al. (2020) for an overview of the formulations. Following the Truncated Anelastic Liquid Approximation (TALA) formulation (Ilangovan et al., 2026), the momentum and mass conservation equations read

$$-\nabla\cdot\boldsymbol{\tau} + \nabla p' = \bar{\rho}\,\bar{\alpha}\,T'\,\hat{\mathbf{r}}, \tag{1}$$

$$\nabla\cdot\big(\bar{\rho}\,\mathbf{u}\big) = 0, \tag{2}$$

with the deviatoric stress tensor

$$\boldsymbol{\tau} \;=\; \eta\big(\nabla\mathbf{u} + \nabla\mathbf{u}^{\top}\big) - \tfrac{2}{3}\,\eta\,(\nabla\cdot\mathbf{u})\,\mathbf{I}, \tag{3}$$

where $\mathbf{u}$ is the velocity, p' the pressure perturbation from the adiabatic hydrostatic state, $\eta = \eta(\mathbf{x}, T)$ the viscosity, which varies spatially and with temperature, $\bar{\rho} = \bar{\rho}(r)$ and $\bar{\alpha} = \bar{\alpha}(r)$ the reference (adiabatic) density and thermal-expansivity profiles, T' the temperature perturbation from the adiabatic reference, and $\hat{\mathbf{r}}$ the radial unit vector.

The energy equation governs the temperature field T:

$$\frac{\partial T}{\partial t} + \mathbf{u}\cdot\nabla T \;-\; \frac{1}{\bar{\rho}\,\bar{c}_p}\,\nabla\cdot\big(k\,\nabla T\big) = \frac{1}{\bar{\rho}\,\bar{c}_p}\,\boldsymbol{\tau}:\dot{\boldsymbol{\varepsilon}} \;+\; \frac{\bar{\alpha}}{\bar{c}_p}\,(\mathbf{u}\cdot\mathbf{g})\,T \;+\; H, \tag{4}$$

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

with $\bar{c}_p = \bar{c}_p(r)$ the reference specific heat, k the thermal conductivity, $\mathbf{g}$ gravity, and the three RHS source terms accounting for shear heating, adiabatic heating, and internal (radiogenic) heating H. The two temperature fields are related by the additive split

$$T(\mathbf{x},t) \;=\; \bar{T}(r) \;+\; T'(\mathbf{x},t), \tag{5}$$

where $\bar{T} = \bar{T}(r)$ is a static radial reference profile, taken either as the analytic steady-conduction solution or as a tabulated geotherm. The system is coupled: (4) is advanced in the full field T, only the deviation T' enters the buoyancy term of (1), and the advection and adiabatic-heating terms in (4) depend on $\mathbf{u}$.

2 Grid and Parallelization

The spherical shell is discretized in two stages. First, the sphere's surface is partitioned into 10 diamonds composed of two spherical triangles based on an icosahedral base grid with 12 vertices (fig. 1a). Each diamond is a curved quadrilateral spanned by 4 icosahedral nodes connected by geodesic arcs. It is refined laterally by recursive midpoint refinement of the underlying spherical triangles (Baumgardner and Frederickson, 1985), yielding $2^\ell \times 2^\ell$ quadrilateral cells per diamond at refinement level ℓ. Second, the surface mesh is extruded radially through N_{rad} layers to fill the shell volume, with flexible spacing (fig. 1). Each hexahedral cell produced by the radial extrusion is split diagonally into two wedge (prism) elements. The splitting is implicit in the compute kernels and not stored in the grid data structure.

For parallel execution, each diamond block is partitioned into subdomains (fig. 1c). Each MPI rank holds one or more subdomains. The unknowns held by an MPI rank are stored in a `Kokkos::View` with dimensions (n_{sd}, n_x, n_y, n_r), where n_{sd} is the number of subdomains, n_x and n_y are the number of cells from refinement in the lateral directions and n_r the number of radial layers. This rank-local data structure exposes a regular 4D index space that is parallelized across GPU threads. At compile time, the View's memory space is bound to the selected Kokkos backend, so the same declaration maps to `cudaMalloc` on NVIDIA GPUs, `hipMalloc` on AMD, or `syclMallocDevice` on Intel, without changing application code. The unknowns are transferred to GPU memory once at initialization and only copied back to the host for I/O. During the solve, all computation stays on the GPU; after each operator apply, boundary DoFs shared between neighbouring subdomains are exchanged using non-blocking, GPU-aware MPI on device-resident send and recv buffers, with packing and atomic-add unpacking implemented as Kokkos kernels in the same backend as the apply itself, avoiding any host staging. Listing 1 shows the corresponding type aliases.

3 Spatial Discretization

The Stokes system (1)–(2) is discretized with the W_1-iso-W_2 / W_1 element pair (velocity W_1-iso-W_2 over pressure W_1), where W_k denotes the wedge element space spanned by the tensor product of a degree-k triangular and a degree-k radial basis, which satisfies the inf-sup (LBB) stability condition: velocity $\boldsymbol{u} \in [W_1]^3$ lives on the refined grid at level ℓ with 6 nodes per wedge (18 DoFs per element), while pressure $p \in W_1$ lives on the coarser grid at level $\ell - 1$. Both use the same tensor-product wedge

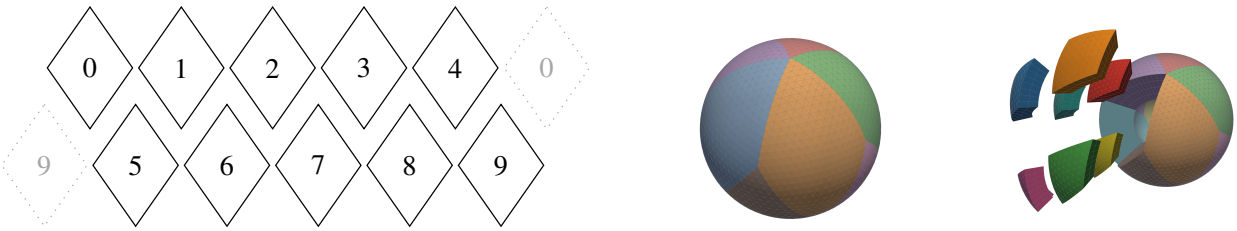
https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

```
using Layout = Kokkos::LayoutRight;

// Scalar grid data (4D: subdomain x lat_x x lat_y x radial)
template <typename ScalarType>
using Grid4DDataScalar = Kokkos::View<ScalarType****, Layout>;

// Vector grid data: structure-of-arrays for coalescing
template <typename ScalarType, int VecDim>
struct Grid4DDataVec {
  Grid4DDataScalar<ScalarType> comp_[VecDim];
};
```

Listing 1. Grid data type aliases. The `ScalarType` template parameter enables mixed-precision instantiation. The `Grid4DDataVec` wrapper stores each vector component in a separate View (structure-of-arrays) for coalesced GPU memory access.



(a) Unfolded diamond layout (b) Assembled shell (c) Exploded view

Figure 1. (a) Unfolded layout of the 10 curvilinear diamonds with IDs 0–9. Ghost diamonds indicate periodic connectivity. (b) 3D Icosahedral spherical shell grid with the 10 diamonds highlighted in colors. (c) Exploded view of a single diamond partitioned into 8 subdomains, each assigned to one MPI rank. An MPI rank can also hold multiple subdomains.

basis functions on different grid levels: on the reference wedge with lateral coordinates $\xi, \eta \geq 0$, $\xi + \eta \leq 1$ and radial coordinate $\zeta \in [-1, 1]$, the six nodal basis functions are the products of P_1 triangle and P_1 radial bases,

$$N_j \;=\; N^{\mathrm{lat}}_{j \bmod 3}(\xi, \eta)\, N^{\mathrm{rad}}_{\lfloor j/3 \rfloor}(\zeta), \qquad j = 0, \ldots, 5, \tag{6}$$

with $N^{\mathrm{lat}}_0 = 1 - \xi - \eta$, $N^{\mathrm{lat}}_1 = \xi$, $N^{\mathrm{lat}}_2 = \eta$ and $N^{\mathrm{rad}}_0 = \frac{1}{2}(1-\zeta)$, $N^{\mathrm{rad}}_1 = \frac{1}{2}(1+\zeta)$, associating nodes 0–2 with the bottom and 3–5 with the top triangular face.

The element integrals are evaluated with a single quadrature point per wedge (two points per hexahedral cell); the accuracy and stability of this choice are analysed in Section 4. The discretization is conforming: nodal degrees of freedom on element boundaries are shared between adjacent wedges. This synergises well with shared-memory tile staging.

As motivated in the introduction, storing the global sparse matrix is infeasible at the target resolution. Instead, the global matrix-vector product $\boldsymbol{y} = A\boldsymbol{x}$ is computed by summing element-local contributions on-the-fly, for example for the viscous

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

part of eq. (1):

$$\boldsymbol{y} = A\,\boldsymbol{x} = \sum_e I_e^\top A_e\, I_e\, \boldsymbol{x}, \tag{7}$$

where I_e is the local-to-global mapping (gather) and A_e is the element matrix arising from the weak form:

$$(A_e)_{ij} = \int_{\Omega_e} \eta \Big[2\,\boldsymbol{\varepsilon}(\boldsymbol{\varphi}_i) : \boldsymbol{\varepsilon}(\boldsymbol{\varphi}_j) - \tfrac{2}{3}\,\mathrm{div}(\boldsymbol{\varphi}_i)\,\mathrm{div}(\boldsymbol{\varphi}_j) \Big]\,\mathrm{d}x, \tag{8}$$

where $\boldsymbol{\varphi}_i$ are the local basis functions. Each element contribution is computed independently: the local DoFs $\boldsymbol{x}_e = I_e\,\boldsymbol{x}$ are gathered from the global source vector, A_e is assembled on-the-fly from geometry and basis functions, the local product $\boldsymbol{y}_e = A_e\,\boldsymbol{x}_e$ is computed, and the result is scattered back via atomic addition ($I_e^\top$). This avoids storing the global matrix entirely, reducing the memory footprint from $\mathcal{O}(N \cdot \mathrm{nnz/row})$ to $\mathcal{O}(N)$ at the cost of recomputing element matrices at every operator application. The details of the kernel implementation, key to achieving high throughput and spatial resolution in our simulations, are explained in the companion performance paper (Böhm et al., 2026c).

4 Single-Point Quadrature and Hourglass Modes

The wedge basis functions (6) span the six-dimensional local space

$$\mathcal{W} \;=\; \mathrm{span}\{\; \underbrace{1}_{\text{constant}} \,;\, \underbrace{\xi,\,\eta,\,\zeta}_{\text{linear}};\; \underbrace{\xi\zeta,\,\eta\zeta}_{\text{mixed bilinear}} \;\}.$$

A single quadrature point at the barycentre $q_c = (\frac{1}{3}, \frac{1}{3}, 0)$ (the centroid–midpoint rule of degree 1 (Felippa, 2004)) evaluates the element integrals (8). Constant and linear basis functions are integrated exactly (centroid rule), only the mixed bilinears are subject to quadrature error. To check for potential velocity modes that are missed by the 1-point quadrature rule, we define the element energy

$$E_e(\boldsymbol{u}) \;=\; V_e\,\eta(q_c)\Big[2\,|\boldsymbol{\varepsilon}(\boldsymbol{u})(q_c)|^2 - \tfrac{2}{3}\,\big(\mathrm{div}\,\boldsymbol{u}(q_c)\big)^2\Big]$$
$$V_e = w_q\,|\det J(q_c)|, \tag{9}$$

with quadrature weight w_q and the Jacobian determinant of the reference map evaluated at q_c. A velocity that is nonzero, but results in a zero element energy, and is therefore invisible to the integral, is called *hourglass mode* in the finite-element literature (Kosloff and Frazier, 1978; Flanagan and Belytschko, 1981).

By (9), a nonzero velocity is an hourglass mode exactly if its gradient vanishes at q_c. Writing $u = a\xi + b\eta + c\zeta + d\xi\zeta + e\eta\zeta \in \mathcal{W}$ per velocity component (omitting constant terms), the condition $\nabla u(q_c) = (a + d\zeta,\; b + e\zeta,\; c + d\xi + e\eta)\big|_{q_c} = 0$ gives $a = b = 0$, $c = -\frac{1}{3}(d+e)$: a two-dimensional null space per component, spanned by

$$w_1 = (\xi - \tfrac{1}{3})\,\zeta, \qquad w_2 = (\eta - \tfrac{1}{3})\,\zeta. \tag{10}$$

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

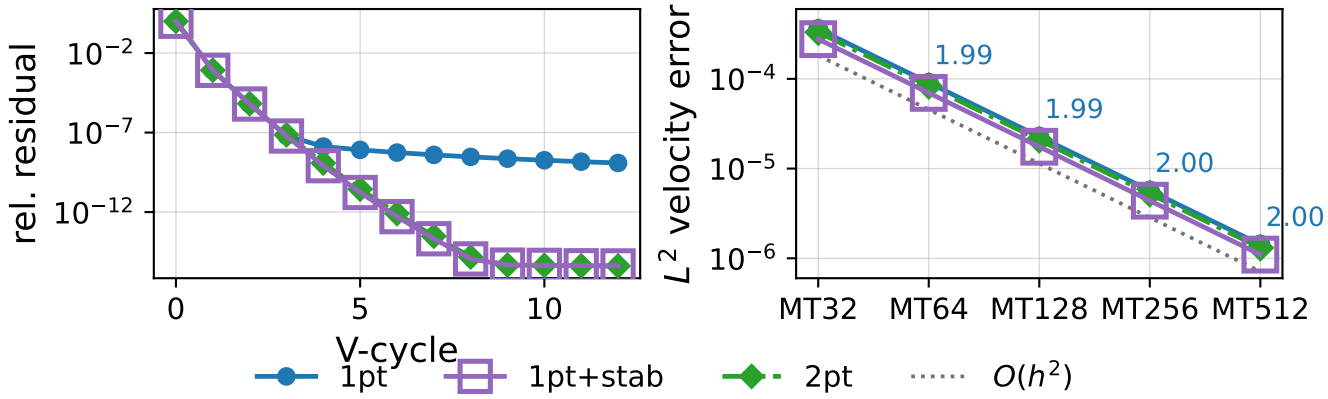


Figure 2. Hourglass control for a smooth analytical solution of the viscous diffusion block of (1) on the spherical shell: stand-alone V-cycle residual on the MT256 mesh (169 M nodes, left) and L^2 velocity discretization error with observed convergence orders over model size MT32–MT512 (right), for one-point quadrature, one-point with hourglass control (11), and the two-radial-point rule. Both remedies recover textbook multigrid convergence (their residual curves coincide, left) at unchanged, optimal-order discretization error.

These are the hourglass modes of the one-point wedge: scalar amplitude profiles carried by a fixed velocity direction, e.g. $\boldsymbol{u} = w_1\, \boldsymbol{e}_\xi$. With three velocity components the element stiffness has $2 \times 3 = 6$ spurious zero-energy modes.

Fortunately, on the whole assembled shell the local modes do not persist, because such local modes can not be stacked together consistently on the whole grid. Thereby, the discretization converges at the optimal rate (Figure 2, right) even with 1 quadrature point. This is contrary to hourglass modes on hexes which persist globally in a checkerboard pattern (Flanagan and Belytschko, 1981). They do, however, degrade stand-alone geometric multigrid, demonstrated by solving the viscous diffusion block of (1) in Figure 2, left. Both a second radial quadrature point and hourglass control (Flanagan and Belytschko, 1981) restore convergence. For the latter, the element matrix is augmented by the rank-6 correction

$$A_e^{\mathrm{stab}} \;=\; A_e \;+\; \varepsilon \sum_{m=1}^{2} \sum_{d=1}^{3} (\boldsymbol{w}_m \otimes \boldsymbol{e}_d)(\boldsymbol{w}_m \otimes \boldsymbol{e}_d)^\top, \tag{11}$$

where $\boldsymbol{w}_m \in \mathbb{R}^6$ is the vector of nodal values of w_m, its coefficient vector in the nodal basis (6), and ε is a small stabilization parameter scaled with the local stiffness. Fused into the gather/scatter phases of the matrix-free kernel (Section 5), it adds a few fused multiply–adds per element and no memory traffic.

5 Matrix-free Assembly Kernel

Element-local assembly is encapsulated in a Kokkos functor (`KOKKOS_FUNCTION operator()`) that is dispatched over the 4D index space via `Kokkos::parallel_for`. The same functor code runs on CPU and GPU backends without modification. Its members are the global source and destination vectors `src_` and `dst_`, held as the structure-of-arrays grid data of Listing 1, and the per-wedge DoF count `nwdofs_`. The host-side `apply()`, the matrix-free matrix-vector multiplication,

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

```
struct ViscousOperator_Textbook {
 Grid4DDataVec<real,3> src_, dst_;
 constexpr int nwdofs_ = 6;   // nodes per wedge

 void apply() {
  Kokkos::parallel_for(
   MDRangePolicy<Rank<4>>({0,..}, {n_sd,nx,ny,nr}), *this);
  Kokkos::fence();
  mpi_halo_exchange(dst_);
 }

 KOKKOS_FUNCTION
 void operator()(int sd, int x, int y, int r) const {
  ...
 }
};
```

Listing 2. Structure of the viscous operator: the element-local assembly is a single Kokkos functor dispatched over the 4D index space (sd, x, y, r), one thread per hexahedral cell.

dispatches the functor over the 4D index space (sd, x, y, r) of subdomain and lateral and radial cell indices, fences, and exchanges halos, so that contributions to nodes on subdomain boundaries are summed across ranks. Listing 2 shows the overall structure.

For each of the two wedges e, the functor evaluates (8) directly against the gathered DoFs $\boldsymbol{x}_e$ without ever forming A_e. At each quadrature point q it contracts the source into the discrete strain rate

$$\boldsymbol{\varepsilon}_q = \sum_{j=1}^{18} \boldsymbol{\varepsilon}(\boldsymbol{\varphi}_j)\big|_q \, (\boldsymbol{x}_e)_j, \tag{12}$$

and accumulates its contraction with every test gradient,

$$(\boldsymbol{y}_e)_i = \sum_q w_q \, |\det J_q| \, \eta_q \times \Big[2\,\boldsymbol{\varepsilon}(\boldsymbol{\varphi}_i)\big|_q : \boldsymbol{\varepsilon}_q - \tfrac{2}{3} \operatorname{div}(\boldsymbol{\varphi}_i)\big|_q \operatorname{tr} \boldsymbol{\varepsilon}_q \Big], \tag{13}$$

for $i = 1, \ldots, 18$. The local result $\boldsymbol{y}_e$ is scattered into the global destination by $I_e^\top$ using atomic addition.

A range of GPU-oriented optimisations turn this textbook baseline into a bandwidth-bound implementation that reaches a substantial fraction of peak performance across NVIDIA, AMD, and Intel GPUs. We defer the description to the companion performance paper (Böhm et al., 2026c).

6 Time Discretization

TERRA-NG provides several schemes for the energy equation (4): the classical streamline-upwind Petrov–Galerkin (SUPG) stabilisation (Brooks and Hughes, 1982), the entropy-viscosity method (Guermond et al., 2011) as used in ASPECT (Kron-

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

bichler et al., 2012), and, as the main scheme, the modified method of characteristics (MMOC) (Douglas and Russell, 1982; Kohl et al., 2022).

A characteristic is the trajectory $\mathbf{X}(t)$ of a fluid particle, $\mathrm{d}\mathbf{X}/\mathrm{d}t = \mathbf{u}(\mathbf{X}, t)$. Following such a particle, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} T\big(\mathbf{X}(t), t\big) \;=\; \frac{\partial T}{\partial t} + \nabla T \cdot \frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} \;=\; \frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T, \tag{14}$$

i.e. the advective derivative in (4) is the ordinary time derivative of the temperature seen by the moving particle. The MMOC exploits this by following the flow backwards: the characteristic that arrives at a node $\mathbf{x}_i$ at t^{n+1} is traced back over one timestep to its departure point

$$\hat{\mathbf{x}}_i \;=\; \mathbf{x}_i - \int_{t^n}^{t^{n+1}} \mathbf{u}\big(\mathbf{X}(t), t\big)\, \mathrm{d}t, \tag{15}$$

integrated with an explicit Runge–Kutta method, and the old temperature is interpolated there, $\hat{T}_i^n = T^n(\hat{\mathbf{x}}_i)$. Replacing the advective derivative by the difference quotient along the characteristic and treating diffusion and sources implicitly on the fixed grid gives

$$\frac{T^{n+1} - \hat{T}^n}{\Delta t} \;-\; \frac{1}{\bar{\rho}\,\bar{c}_p} \nabla \cdot \big(k \nabla T^{n+1}\big) \;=\; Q, \tag{16}$$

with Q the right-hand side of (4). Each timestep, thus, splits into an advection step, i.e. the evaluation of T^n at all departure points, and a diffusion step, the solution of (16) with matrix-free preconditioned CG. As advection is carried along the flow rather than through a stencil, the scheme is not CFL-bound for stability, admits considerably larger timesteps and introduces little numerical diffusion. In TERRA-NG, each particle is fully virtual (holding no information) and traced back to its departure point by a dedicated thread in a Kokkos kernel. At its departure point, the temperature is interpolated using linear, or higher order interpolation, with cross GPU boundaries handled by enlarged boundary layers. A more detailed analysis of the scheme can be found in (Kohl et al., 2022).

7 Solver Structure

The Stokes saddle-point matrix is preconditioned by the standard block-triangular Schur-complement approach (Elman et al., 2005) widely-used across mantle convection codes (May and Moresi, 2008; Burstedde et al., 2013; Kronbichler et al., 2012). Writing the discretised system as $\left(\begin{smallmatrix} A & B^T \\ B & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} \mathbf{u} \\ p \end{smallmatrix}\right) = \left(\begin{smallmatrix} \mathbf{f} \\ \mathbf{g} \end{smallmatrix}\right)$, the preconditioner reads

$$\mathcal{P}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1} B^T \hat{S}^{-1} \\ 0 & \hat{S}^{-1} \end{pmatrix}, \tag{17}$$

with $\hat{S}^{-1}$ approximating the inverse of the Schur complement $S = -BA^{-1}B^T$ by the scaled diagonal of the inverse-viscosity-weighted pressure mass matrix,

$$\hat{S}^{-1} = -\operatorname{diag}\big(M_{1/\eta}\big)^{-1}, \qquad \big(M_{1/\eta}\big)_{ij} = \int_\Omega \frac{1}{\eta}\, \psi_i\, \psi_j \, \mathrm{d}\Omega, \tag{18}$$

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

with the pressure basis functions ψ_i: for moderate viscosity contrasts, $M_{1/\eta}$ is spectrally equivalent to the (negative) Schur complement for variable viscosity (Elman et al., 2005). The velocity-block inverse A^{-1} is realised approximately by a matrix-free geometric multigrid V-cycle with Chebyshev smoothing (Elman et al., 2005; Rudi et al., 2015), which is the dominant cost of every FGMRES (Saad, 1993) iteration in the outer solver.

The V-cycle applies two pre- and two post-smoothing steps of a second-degree Chebyshev polynomial smoother on every level and solves the coarsest-grid system with conjugate gradients. The radial viscosity structure of the mantle is what limits the performance of these solver components: realistic profiles vary by orders of magnitude over a few hundred kilometres, with the steepest gradients at the base of the lithosphere and across the asthenosphere (Lin et al., 2022; Stotz et al., 2018). Such coefficient contrasts degrade the smoother and coarse-grid correction of the velocity multigrid as well as the Schur approximation (18). We therefore consider two radial profiles of different severity: the profile of Lin et al. 2022, spanning three orders of magnitude (4.7×10^{20} – 5×10^{23} Pa s), and the profile of Stotz et al. 2018, spanning four orders (5.8×10^{19} – 7×10^{23} Pa s) with a pronounced low-viscosity asthenosphere.

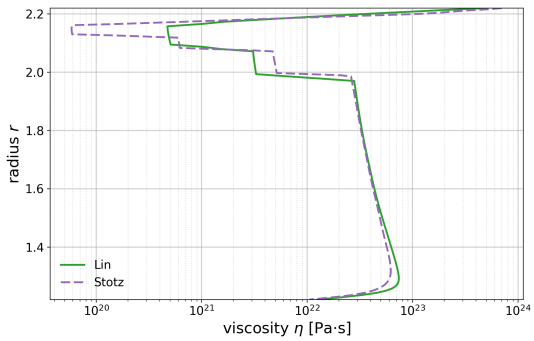
Figure 3 documents convergence of the full Stokes solve for both profiles. The left table reports the number of preconditioned FGMRES iterations needed to reduce the relative residual by six orders of magnitude; the right table reports the relative residual actually achieved after a fixed budget of 10 outer FGMRES iterations, which is a realistic configuration in production runs. The results confirm that the multigrid preconditioner delivers mesh-independent convergence with real-world viscosity profiles. There is a light drift for the large models, which is expected for the light-weight pressure-mass preconditioner (in contrast to a stronger BFBT-type preconditioner (Rudi et al., 2017; Burkhart et al., 2026)), but convergence is satisfactory. The cell-to-cell coefficient contrast seen by the smoother and the coarse grids decreases with model size, as the smooth profile is resolved more finely, so the solve initially gets easier as the problem size grows.

For even sharper contrasts, TERRA-NG possesses a GPU-accelerated, wedge-based implementation of adaptive Galerkin coarsening (Böhm et al., 2026b). There, assembled Galerkin coarse-grid operators are built element-wise only where the viscosity gradient exceeds a threshold, typically a thin band along the boundary layers, while the cheap matrix-free re-discretisation is kept everywhere else, restoring robust multigrid convergence under strong coefficient contrasts at minimal memory overhead.

8 Verification Benchmarks

We verify the coupled Stokes-energy framework against a set of spherical shell convection solutions, explored early on for steady flow (e.g. Bercovici et al. (1989); Ratcliff et al. (1996); Zhong et al. (2000); Richards et al. (2001)), extended in the mantle convection benchmark suite of Zhong et al. (2008), and compiled and reused in the recent code intercomparison of Euen et al. (2023) and the HyTeG validation of Ilangovan et al. (2026). We pick cases that span the relevant parameter range, from mild low-Rayleigh tetrahedral convection (A1–A7, increasing viscosity contrast) up to vigorous high-Rayleigh cubic convection at $\mathrm{Ra} = 10^5 - 10^7$ with C1–C4:

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.



MT	Lin	Stotz
MT32	41	49
MT64	32	45
MT128	29	37
MT256	26	33
MT512	27	33
MT1024	31	40
MT2048	31	42

FGMRES iters to 10^{-6} relative residual.

MT	Lin	Stotz
MT32	3.7×10^{-4}	3.5×10^{-4}
MT64	2.9×10^{-4}	3.0×10^{-4}
MT128	2.4×10^{-4}	3.1×10^{-4}
MT256	1.63×10^{-4}	2.37×10^{-4}
MT512	1.12×10^{-4}	2.25×10^{-4}
MT1024	6.88×10^{-4}	1.57×10^{-3}
MT2048	7.09×10^{-4}	1.52×10^{-3}

Relative residual after 10 FGMRES iters

Figure 3. Left: radial viscosity profiles of Lin et al. (2022) and Stotz et al. (2018) used for the convergence study. Centre: full Stokes FGMRES iterations to reduce the relative residual by six orders of magnitude. Right: relative residual reached after a fixed budget of 10 outer FGMRES iterations.

Table 2. Steady-state $\mathrm{Nu_{top}}$ vs. the reference intervals from Euen et al. (2023); Ilangovan et al. (2026) (A3, C1, C3), the CitcomS reference values of Zhong et al. (2008) (A1, A4–A7, C4), and the isoviscous Ra=10^7 value of Garrido-Tomasini et al. (2025) (C1* denotes C1 at Ra=10^7).

case	A1	A3	A4	A5	A6	A7	C1	C3	C4	C1*
$\mathrm{Nu_{top}}$ (TERRA-NG)	3.49	3.162	2.99	2.54	2.055	2.716	7.714	6.778	6.496	31.1
reference	3.5126	[3.140, 3.190]	2.9354	2.5468	2.0414	2.7382	[7.370, 7.810]	[6.500, 6.790]	6.4800	31.91

case	Ra	$\Delta\eta$	T perturb
A series	7×10^3	$\{1, 20, 10^2, 10^3, 10^4, 10^5\}$	Y_3^2
C series	1×10^5	$\{1, 30, 100\}$	$Y_4^0 + \frac{5}{7}Y_4^4$
C1 (Ra=10^7)	1×10^7	1	$Y_4^0 + \frac{5}{7}Y_4^4$

We run all cases with an MT256 resolution, use the non-dimensional shell $r \in [1.22, 2.22]$, free-slip boundary conditions on the top and bottom surface, isothermal Dirichlet temperature conditions ($T_{\mathrm{cmb}} = 1$, $T_{\mathrm{surf}} = 0$), and the Frank–Kamenetskii viscosity law $\eta(T) = \exp\big(-E(T - \frac{1}{2})\big)$, with the viscosity contrast $\Delta\eta = \exp(E)$ between top and bottom matching the value tabulated by Zhong et al. (2008). Each case is advanced until the radially averaged temperature profile and the surface Nusselt number $\mathrm{Nu_{top}}$ stabilise; we then report the steady-state $\langle T\rangle(r)$ profile (Figure 4) and the steady-state $\mathrm{Nu_{top}}$. For A3, C1, and C3 the computed $\mathrm{Nu_{top}}$ lies inside the reference interval reported by the intercomparison studies; for the temperature-dependent-viscosity cases A4 and A5 we compare against the CitcomS reference values of Zhong et al. (2008); and for the high-Ra C1 ($\mathrm{Ra} = 10^7$) case we compare against the recent isoviscous spherical-shell result of Garrido-Tomasini et al. (2025). Across the board, the Nusselt number lies in the expected interval or is very close to the reference value given by literature.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

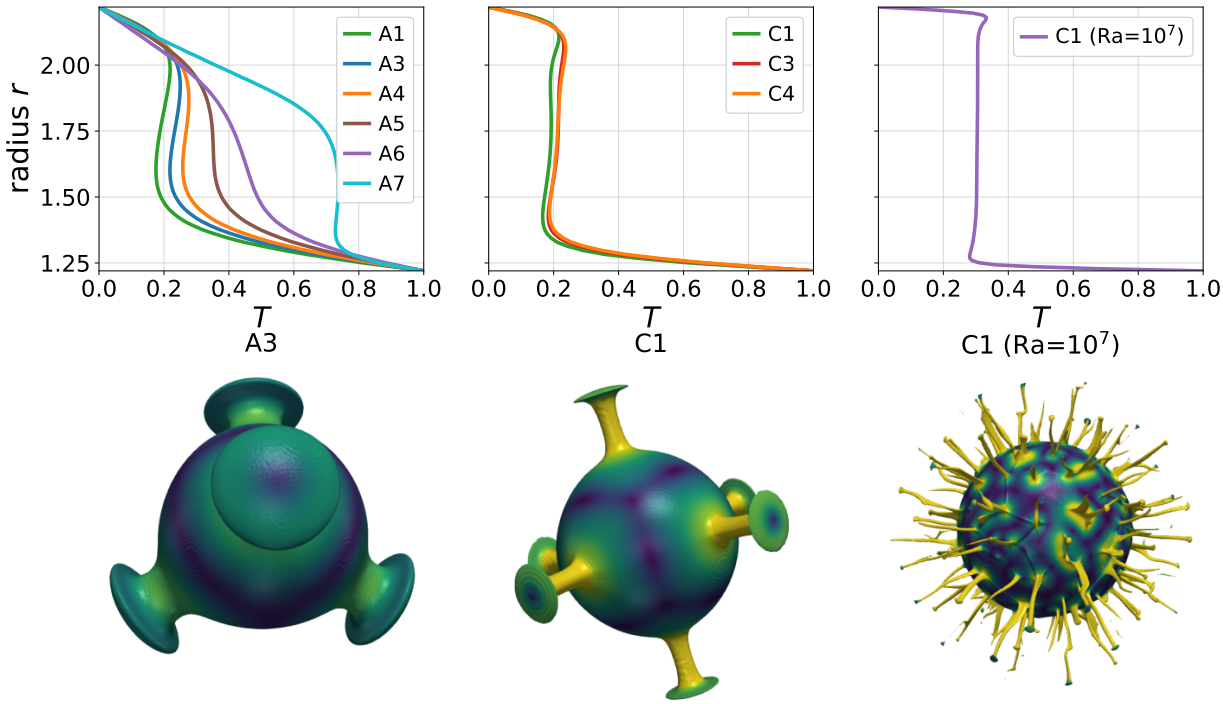


Figure 4. Verification benchmarks: steady-state radial temperature profiles $\langle T \rangle(r)$ (top row) and, beneath each, the steady-state $T = 0.5$ isosurface of the named case, colored by velocity magnitude at the same time. These benchmarks were run using MMOC for the energy equation. Entropy viscosity gave the same or very similar results in terms of final isosurface, Nusselt number and radial profile.

9 Strong and Weak Scaling of TERRA-NG

The simulations in this section advance global mantle-circulation modelling to resolutions that, to the best of our knowledge, have not been reported before: coupled Stokes–energy timestepping with compressibility and variable viscosity beyond MT512, which is the upper end of published global models, up to MT4096 and, in a low-memory mode, MT8192. Furthermore, we show portability of the code to five supercomputers spanning three GPU vendors. Table 3 lists the five host systems on which the simulations are run. We included 2 NVIDIA, 2 AMD and 1 Intel machine for a representative vendor study. We used benchmarking allocations through the EuroHPC Joint Undertaking for most machines, except LUMI-G, where we computed using an extreme-scale access.

We evaluate the feasibility of running coupled Stokes–energy mantle-convection simulations up to a given spatial resolution on each of the considered GPU supercomputers, and identify the optimal node configuration for every resolution through a per-MT-level strong-scaling sweep. Each timestep solves the Stokes saddle-point system with 10 preconditioned FGMRES iterations and the energy equation with 50 preconditioned FGMRES iterations. We run 10 timesteps with noslip-freeslip bound-

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

Table 3. Host supercomputers of the feasibility study, with operating compute centre, accelerator-partition node count, per-node GPU architecture, and aggregate main memory of the GPU partition (host DRAM + per-accelerator HBM). All systems carry 512 GB of node-level main memory (DDR4/DDR5 on JUWELS Booster, MareNostrum 5, LUMI-G and SuperMUC-NG; unified HBM3 on the Hunter MI300A APU nodes).

system	vendor	computing centre	nodes	Architecture	main memory
JUWELS Booster	NVIDIA	JSC, Jülich, DE	936	A100 SXM4	480 TB
MareNostrum 5 ACC	NVIDIA	BSC, Barcelona, ES	1,120	H100 SXM5	573 TB
LUMI-G	AMD	CSC, Kajaani, FI	2,978	MI250X	1.5 PB
Hunter	AMD	HLRS, Stuttgart, DE	188	4 MI300A APU	96 TB
SuperMUC-NG Phase 2	Intel	LRZ, Munich, DE	234	PVC	120 TB

ary conditions and the Lin viscosity profile. The low-memory mode sets the FGMRES restarts to 5 instead of 10 and stores the Krylov basis vectors in single precision, reducing the memory the Krylov basis requires.

Figure 5 plots the average time per simulation timestep over the MT configurations. Each line is a strong-scaling sweep: the model size is held fixed and spread over more and more GPUs, so the curve falls until the per-device work no longer pays for the added communication, and speed-ups from more devices diminish. The figure also carries a weak-scaling reading across its lines. Every step to the next MT level multiplies the number of unknowns by eight, four times laterally and twice radially, so pairing that step with an eight-fold increase in devices keeps the work per device constant. Read along that diagonal, the curve starts line up on a roughly straight weak-scaling line, especially visible for LUMI-G: MT256 on 2 GCDs, MT512 on 16, MT1024 on 128, MT2048 on 1024 and MT4096 on 8192 all hold 126 M DoFs per GCD, and the time per timestep grows from 26.2 to 30.9 s across that chain, an 18 % increase for a 4096-fold growth in both device count and problem size.

We observe that simulations up to MT128 run with $<$ 10 seconds per timestep on all systems. Resolutions of MT256 – MT1024, at and beyond the upper end of published global mantle-convection simulations, run routinely on most systems with $<$ 20 seconds per timestep. TERRA-NG allows running MT4096 (1.4×10^{12} DoFs) reaching the 1-km-per-gridpoint resolution with the extreme-scale access on LUMI-G at $\sim$ 30 seconds per timestep. On the full machine, an MT8192 with 1.1×10^{13} DoFs is feasible in low-memory-mode which reaches sub-1-km-resolution. Already MT1024 exceeds, to the best of the authors' knowledge, any published global mantle-convection simulation with compressibility and variable viscosity, MT4096 and MT8192 push this frontier by one to two further orders of magnitude in DoFs.

10 End-to-end Simulation

To demonstrate TERRA-NG as a production tool, we run an end-to-end mantle-circulation model at MT1024 on 64 nodes (512 PVC tiles, one rank per tile) of SuperMUC-NG Phase 2: at 5.6 km radial and $\sim$7 km lateral spacing, i.e. 1.7×10^{10} velocity–pressure DoF per Stokes solve, and 2.2×10^{10} DoF in total once the temperature field is counted. The model is TALA-compressible with a $\bar{\rho}$-weighted buoyancy on an adiabatic reference state, free-slip at both boundaries, and fixed temperatures

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

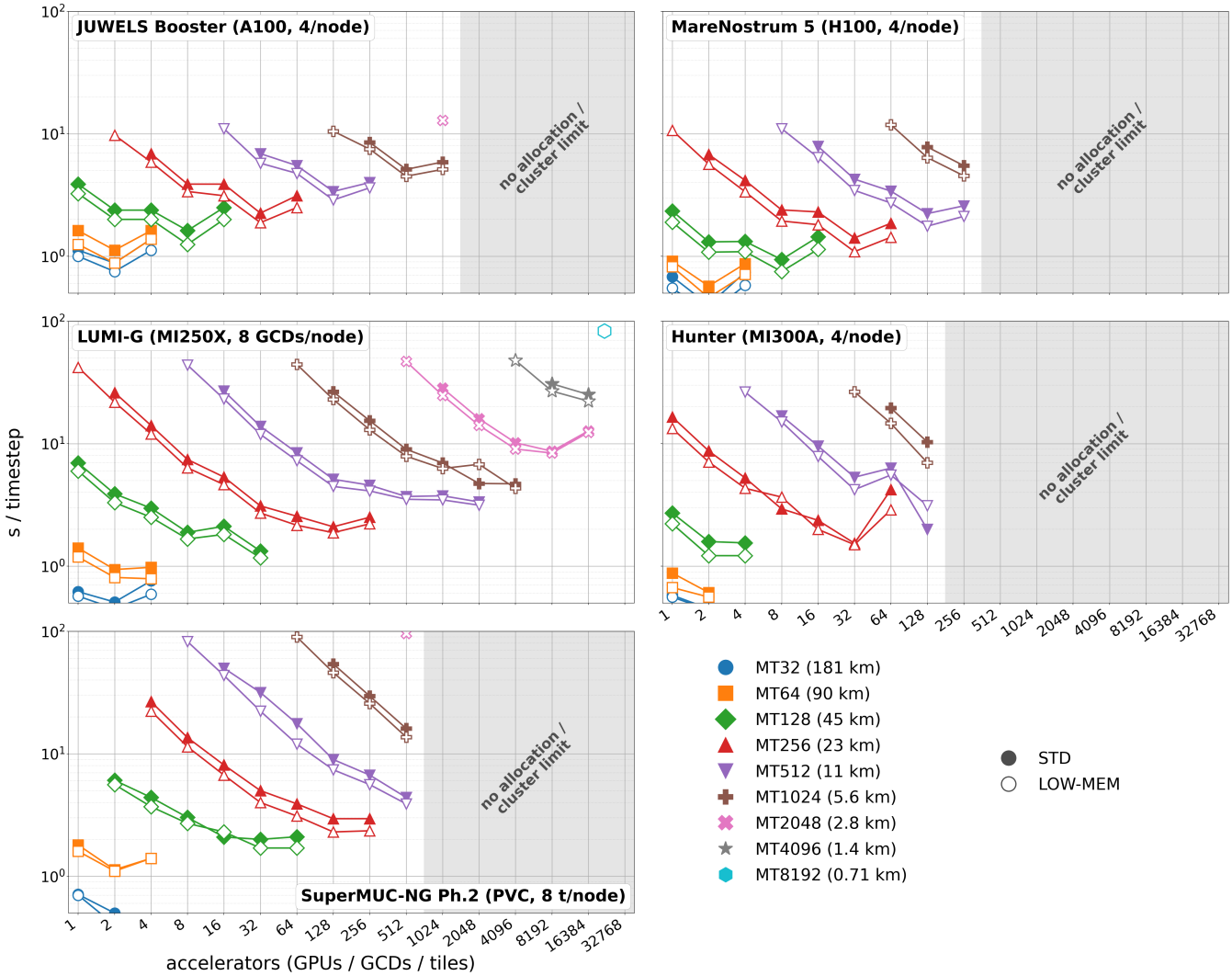


Figure 5. Strong scaling of the coupled Stokes–energy timestep across five GPU supercomputers, with one line per MT Model and corresponding grid spacing (Table 1).The scaling results were run using entropy viscosity for the energy equation, because MMOC was implemented after the accesses to some of these machines.

of 4200 K at the CMB and 300 K at the surface. The viscosity is the radial profile of Lin et al. (2022) multiplied by a Frank–Kamenetskii temperature dependence of contrast $\Delta\eta_T = 100$. The reference viscosity is 3.4×10^{20} Pa s, the minimum of that radial profile, which puts the Rayleigh number formed with it at $\mathrm{Ra} \approx 5 \times 10^8$. The initial condition is a conductive profile perturbed by a broadband spherical-harmonic field of degrees $\ell = 8$–96 (amplitude 0.05), so that no low-degree symmetry is imposed.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

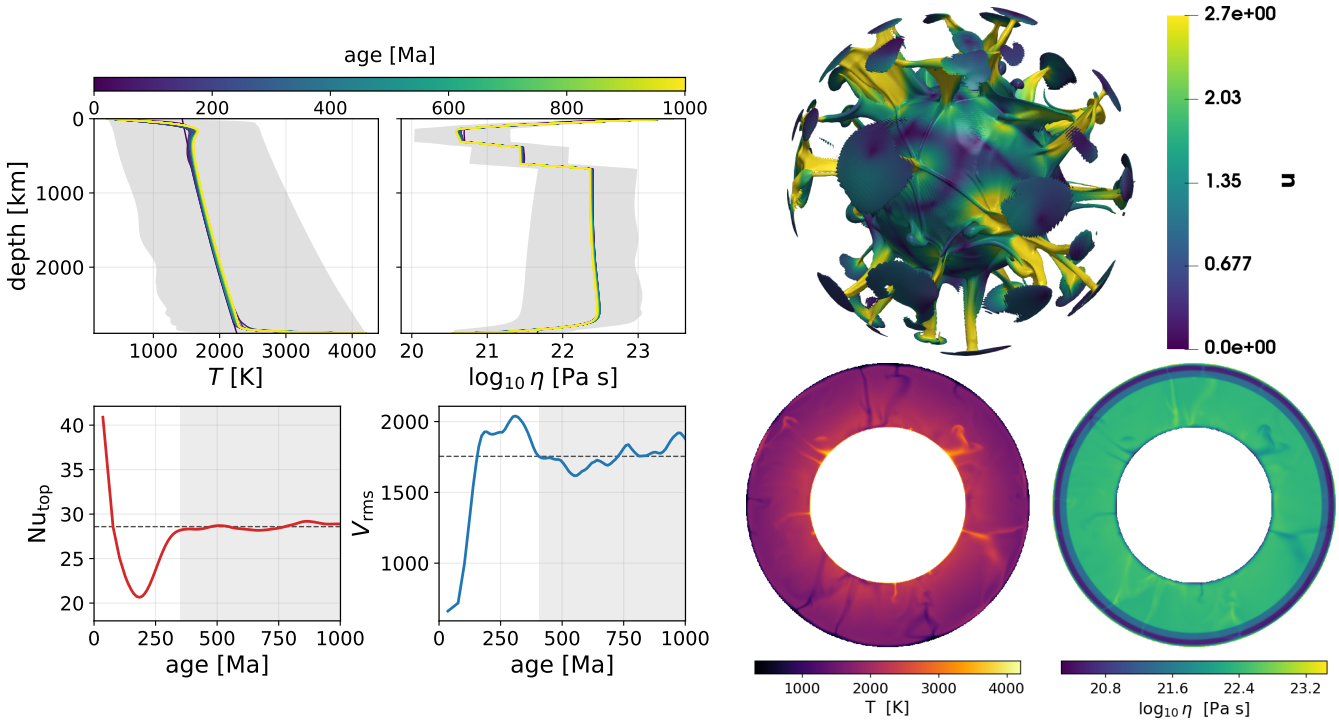


Figure 6. End-to-end MT1024 simulation. Left block, top row: radial-mean temperature and viscosity every 100 timesteps, coloured by model age, over the lateral min–max range of the final state in grey. The imposed Lin staircase is overlaid by the Frank–Kamenetskii response of the lid, asthenosphere and basal layer to the thermal boundary layers. Bottom row: surface Nusselt number and root-mean-square velocity against model age, with the steady window shaded and its mean dashed. Right, top: isosurface of the temperature deviation from the radial mean at $+400$ K, coloured by velocity magnitude. Right, bottom: meridional cross-cut of temperature and viscosity through the same state at $t = 1.00$ Gyr. This end-to-end simulation was run with MMOC for the energy equation.

Each timestep solves the Stokes system with 10 FGMRES iterations on the multigrid-preconditioned velocity block and transports the temperature with the MMOC semi-Lagrangian scheme at a Courant number of 2.7, whose diffusion step takes at most 20 Krylov iterations at a 10^{-3} relative tolerance, at $\sim$19 s of wall time per step including output. The CFL-controlled timestep shrinks from 0.32 Ma in the initial layered state to 0.10 Ma once the convective flow is fully developed. 8800 steps advance the model over 1.00 Gyr across two chained 24-hour allocations, i.e. $\sim$23 Ma of mantle evolution per compute hour. The two diagnostics settle at different times (Figure 6, bottom left): the surface Nusselt number is flat at $\mathrm{Nu}_{\mathrm{top}} \sim 28.6$ from $\sim$350 Ma onward, whereas the root-mean-square velocity overshoots and only then settles at $V_{\mathrm{rms}} \sim 1.8 \times 10^3$.

Conclusions

Global mantle-circulation modelling has long been resolution-bound. TERRA-NG moves that bound to kilometre-scale grid spacing through a deliberately narrow design: a single application, one mesh geometry, and discretization tailored towards that

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

geometry rather than a general finite-element abstraction. That narrow design is what makes the substantial domain-specific optimizations possible, and it is those optimizations, together with the hardware capabilities of modern GPUs, that carry the resolution. Thanks to Kokkos, TERRA-NG runs on NVIDIA, AMD and Intel GPUs, demonstrated on five supercomputers, enabling compressible, variable-viscosity mantle circulation at

- ~ 5.6 km grid spacing (22 B DoFs) in under 20 seconds per timestep on all five systems,
- ~ 2.8 km and ~ 1.4 km grid spacing (177 B and 1.4 T DoFs) on the extreme-scale allocation of LUMI-G, the latter at ~ 30 seconds per timestep.

The value of these resolutions lies less in the individual model than in what becomes routine. Mantle-circulation models supply quantitative input to neighbouring geoscience disciplines, the core–mantle-boundary heat-flux pattern that drives geodynamo simulations, and the mantle heterogeneity structure against which seismological, geodetic and geologic observations can be interpreted technically, but only if they can be produced at sufficient resolution. Timesteps of seconds at kilometre-scale resolution and high model complexity turn such campaigns from single hero runs into a production workflow, which is where the results of this paper are meant to lead.

AI disclosure. TERRA-NG contains a subset of code that was developed with Claude Code (Anthropic) as an interactive coding assistant, under the close supervision of the authors, who gave detailed instructions and requirements. All functionality, whether written by humans or with machine assistance, has been thoroughly reviewed and tested by the authors, who take full responsibility for it. During the preparation of this manuscript, the authors used Claude (Anthropic) to draft and improve the readability and language of the text, while all structure and content were provided by the authors. No technical claims, figures or interpretations are AI-generated. The authors reviewed and edited all content and take full responsibility for it.

Code and data availability. TERRA-NG is developed openly at https://github.com/mantleconvection/TERRA-NG under the GNU General Public License v3. The exact version described in this paper, including the configurations and batch scripts under `apps/mantlecirculation/terrangpaper/` that reproduce all figures and tables, is archived on Zenodo at https://doi.org/10.5281/zenodo.22938125 (Böhm et al., 2026a).

Author contributions. FB designed and implemented the matrix-free wedge operators, carried out the verification benchmarks and the scaling campaign, and wrote the manuscript. NK designed the grid and parallelisation layer, the linear-algebra and solver infrastructure and the test framework, and contributed to the operator implementation. GR and FB developed the mantle-circulation application and its diagnostics. PI contributed finite-element operators and linear forms. PI and FB contributed the MMOC scheme. FR and GR contributed the plate velocity boundary conditions in the mantle-circulation application. GR, PI, MM, BSAS and HPB provided the geodynamical problem setting, defined the verification benchmarks and interpreted the results. HK and UR supervised the HPC design and the performance work. FB, BSAS, HK, HPB and UR acquired funding and compute resources. All authors read, revised and approved the manuscript.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich–Alexander–Universität Erlangen–Nürnberg (FAU). The hardware is partially funded by the German Research Foundation (DFG). We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Regular Access call (EHPC-BEN-2026B05-001). We acknowledge the EuroHPC Joint Undertaking for awarding us access to MareNostrum5 at BSC, Spain (EHPC-BEN-2026B04-050). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer SuperMUC-NG at Leibniz Supercomputing Centre (www.lrz.de) Project pn25xe. The simulations were performed on the national supercomputer HPE Cray EX4000 Hunter at the High Performance Computing Center Stuttgart (HLRS).

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

References

Bauer, S., Huber, M., Ghelichkhan, S., Mohr, M., Rüde, U., and Wohlmuth, B.: Large-scale simulation of mantle convection based on a new matrix-free approach, Journal of Computational Science, 31, 60–76, https://doi.org/10.1016/j.jocs.2018.12.006, 2019.

Bauer, S., Bunge, H.-P., Drzisga, D., Ghelichkhan, S., Huber, M., Kohl, N., Mohr, M., Rüde, U., Thönnes, D., and Wohlmuth, B.: TerraNeo – Mantle Convection Beyond a Trillion Degrees of Freedom, in: Software for Exascale Computing – SPPEXA 2016–2019, vol. 136 of *Lecture Notes in Computational Science and Engineering*, pp. 569–610, Springer, https://doi.org/10.1007/978-3-030-47956-5_19, 2020.

Baumgardner, J. R.: Three-dimensional treatment of convective flow in the Earth's mantle, Journal of Statistical Physics, 39, 501–511, https://doi.org/10.1007/BF01008348, 1985.

Baumgardner, J. R. and Frederickson, P. O.: Icosahedral discretization of the two-sphere, SIAM Journal on Numerical Analysis, 22, 1107–1115, https://doi.org/10.1137/0722066, 1985.

Bercovici, D., Schubert, G., Glatzmaier, G. A., and Zebib, A.: Three-dimensional thermal convection in a spherical shell, Journal of Fluid Mechanics, 206, 75–104, https://doi.org/10.1017/S0022112089002235, 1989.

Böhm, F., Kohl, N., Ilangovan, P., Robl, G., Rezaei, F., Mohr, M., Schuberth, B. S. A., Köstler, H., Bunge, H.-P., and Rüde, U.: TERRA-NG v1.0: source code and reproduction inputs, https://doi.org/10.5281/zenodo.22938125, 2026a.

Böhm, F., Kohl, N., Köstler, H., and Rüde, U.: Large-scale Multigrid with Adaptive Galerkin Coarsening, https://doi.org/10.48550/arXiv.2511.13109, arXiv preprint arXiv:2511.13109, accepted for publication in SIAM Journal on Scientific Computing, 2026b.

Böhm, F., Kohl, N., Köstler, H., and Rüde, U.: Performance Analysis of Low-Order, GPU-accelerated Finite Element Kernels using Kokkos, https://doi.org/10.48550/arXiv.2609.21681, arXiv preprint arXiv:2609.21681, 2026c.

Brooks, A. N. and Hughes, T. J. R.: Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations, Computer Methods in Applied Mechanics and Engineering, 32, 199–259, https://doi.org/10.1016/0045-7825(82)90071-8, 1982.

Bunge, H.-P. and Baumgardner, J. R.: Mantle convection modeling on parallel virtual machines, Computers in Physics, 9, 207–215, https://doi.org/10.1063/1.168525, 1995.

Bunge, H.-P., Richards, M. A., and Baumgardner, J. R.: Effect of depth-dependent viscosity on the planform of mantle convection, Nature, 379, 436–438, https://doi.org/10.1038/379436a0, 1996.

Bunge, H.-P., Richards, M. A., and Baumgardner, J. R.: A sensitivity study of three-dimensional spherical mantle convection at 10^8 Rayleigh number: Effects of depth-dependent viscosity, heating mode, and an endothermic phase change, Journal of Geophysical Research: Solid Earth, 102, 11 991–12 007, https://doi.org/10.1029/96JB03806, 1997.

Burkhart, A., Kohl, N., Wohlmuth, B., and Zawallich, J.: A robust matrix-free approach for large-scale non-isothermal high-contrast viscosity Stokes flow on blended domains with applications to geophysics, GEM - International Journal on Geomathematics, 17, https://doi.org/10.1007/s13137-025-00280-5, 2026.

Burstedde, C., Stadler, G., Alisic, L., Wilcox, L. C., Tan, E., Gurnis, M., and Ghattas, O.: Large-scale adaptive mantle convection simulation, Geophysical Journal International, 192, 889–906, https://doi.org/10.1093/gji/ggs070, 2013.

Clevenger, T. C. and Heister, T.: Comparison between algebraic and matrix-free geometric multigrid for a Stokes problem on adaptive meshes with variable viscosity, Numerical Linear Algebra with Applications, 28, e2375, https://doi.org/10.1002/nla.2375, 2021.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

Davies, D. R., Wilson, C. R., and Kramer, S. C.: Fluidity: A fully unstructured anisotropic adaptive mesh computational modeling framework for geodynamics, Geochemistry, Geophysics, Geosystems, 12, Q06 001, https://doi.org/10.1029/2011GC003551, 2011.

Davies, D. R., Davies, J. H., Bollada, P. C., Hassan, O., Morgan, K., and Nithiarasu, P.: A hierarchical mesh refinement technique for global 3-D spherical mantle convection modelling, Geoscientific Model Development, 6, 1095–1107, https://doi.org/10.5194/gmd-6-1095-2013, 2013.

Davies, D. R., Kramer, S. C., Ghelichkhan, S., and Gibson, A.: Towards automatic finite-element methods for geodynamics via Firedrake, Geoscientific Model Development, 15, 5127–5166, https://doi.org/10.5194/gmd-15-5127-2022, 2022.

Davies, G. F.: Dynamic Earth, https://doi.org/10.2277/0521590671, 1999.

Douglas, Jr., J. and Russell, T. F.: Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures, SIAM Journal on Numerical Analysis, 19, 871–885, https://doi.org/10.1137/0719063, 1982.

Elman, H. C., Silvester, D. J., and Wathen, A. J.: Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics, Oxford University Press, https://doi.org/10.1093/oso/9780198528678.001.0001, 2005.

Euen, G. T., Liu, S., Gassmöller, R., Heister, T., and King, S. D.: A comparison of 3-D spherical shell thermal convection results at low to moderate Rayleigh number using ASPECT (version 2.2.0) and CitcomS (version 3.3.1), Geoscientific Model Development, 16, 3221–3239, https://doi.org/10.5194/gmd-16-3221-2023, 2023.

Felippa, C. A.: A compendium of FEM integration formulas for symbolic work, Engineering Computations, 21, 867–890, https://doi.org/10.1108/02644400410554362, 2004.

Flanagan, D. P. and Belytschko, T.: A uniform strain hexahedron and quadrilateral with orthogonal hourglass control, International Journal for Numerical Methods in Engineering, 17, 679–706, https://doi.org/10.1002/nme.1620170504, 1981.

Garrido-Tomasini, J., Lenardic, A., Moore, W., and Weller, M.: Infinite Prandtl number, mixed-heated convection in a spherical domain: velocity and heat flux scaling trends, Geophysical Journal International, 241, 203–218, https://doi.org/10.1093/gji/ggaf039, 2025.

Gassmöller, R., Dannberg, J., Bangerth, W., Heister, T., and Myhill, R.: On formulations of compressible mantle convection, Geophysical Journal International, 221, 1264–1280, https://doi.org/10.1093/gji/ggaa078, 2020.

Glatzmaier, G. A.: Numerical simulations of mantle convection: time-dependent, three-dimensional, compressible, spherical shell, Geophysical & Astrophysical Fluid Dynamics, 43, 223–264, https://doi.org/10.1080/03091928808213626, 1988.

Gough, D. O.: The Anelastic Approximation for Thermal Convection, Journal of the Atmospheric Sciences, 26, 448–456, https://doi.org/10.1175/1520-0469(1969)026<0448:TAAFTC>2.0.CO;2, 1969.

Guermond, J.-L., Pasquetti, R., and Popov, B.: Entropy viscosity method for nonlinear conservation laws, Journal of Computational Physics, 230, 4248–4267, https://doi.org/10.1016/j.jcp.2010.11.043, 2011.

Heister, T., Dannberg, J., Gassmöller, R., and Bangerth, W.: High accuracy mantle convection simulation through modern numerical methods – II: realistic models and problems, Geophysical Journal International, 210, 833–851, https://doi.org/10.1093/gji/ggx195, 2017.

Ilangovan, S. P., Kohl, N., and Mohr, M.: Highly scalable geodynamic simulations with HyTeG, Geoscientific Model Development, 19, 1455–, https://doi.org/10.5194/gmd-19-1455-2026, 2026.

Jarvis, G. T. and McKenzie, D. P.: Convection in a compressible fluid with infinite Prandtl number, Journal of Fluid Mechanics, 96, 515–583, https://doi.org/10.1017/S002211208000225X, 1980.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

Kohl, N., Thönnes, D., Drzisga, D., Bartuschat, D., and Rüde, U.: The HyTeG finite-element software framework for scalable multigrid solvers, International Journal of Parallel, Emergent and Distributed Systems, 34, 477–496, https://doi.org/10.1080/17445760.2018.1506453, 2019.

Kohl, N., Mohr, M., Eibl, S., and Rüde, U.: A massively parallel Eulerian–Lagrangian method for advection-dominated transport in viscous fluids, SIAM Journal on Scientific Computing, 44, C260–C285, https://doi.org/10.1137/21M1402510, 2022.

Kosloff, D. and Frazier, G. A.: Treatment of hourglass patterns in low order finite element codes, International Journal for Numerical and Analytical Methods in Geomechanics, 2, 57–72, https://doi.org/10.1002/nag.1610020105, 1978.

Kronbichler, M. and Kormann, K.: A generic interface for parallel cell-based finite element operator application, Computers and Fluids, 63, 135–147, https://doi.org/10.1016/j.compfluid.2012.04.012, 2012.

Kronbichler, M. and Ljungkvist, K.: Multigrid for matrix-free high-order finite element computations on graphics processors, ACM Transactions on Parallel Computing, 6, 2:1–2:32, https://doi.org/10.1145/3322813, 2019.

Kronbichler, M., Heister, T., and Bangerth, W.: High accuracy mantle convection simulation through modern numerical methods, Geophysical Journal International, 191, 12–29, https://doi.org/10.1111/j.1365-246X.2012.05609.x, 2012.

Lee, K.-M., Jang, D.-K., Thieulot, C., and So, B.-D.: High-Performance Geodynamics on GPUs Using the PETSc CUDA Backend (GAUZZ v1.0.0), EGUsphere, 2026, 1–39, https://doi.org/10.5194/egusphere-2026-2528, 2026.

Lin, Y.-A., Colli, L., Wu, J., and Schuberth, B. S. A.: Where are the proto-South China Sea slabs? SE Asian plate tectonics and mantle flow history from global mantle convection modeling, Journal of Geophysical Research: Solid Earth, 127, e2022JB024 633, https://doi.org/10.1029/2022JB024633, 2022.

May, D. A. and Moresi, L.: Preconditioned iterative methods for Stokes flow problems arising in computational geodynamics, Physics of the Earth and Planetary Interiors, 171, 33–47, https://doi.org/10.1016/j.pepi.2008.07.036, 2008.

May, D. A., Brown, J., and Le Pourhiet, L.: pTatin3D: High-Performance Methods for Long-Term Lithospheric Dynamics, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC14), pp. 274–284, https://doi.org/10.1109/SC.2014.28, 2014.

May, D. A., Brown, J., and Le Pourhiet, L.: A scalable, matrix-free multigrid preconditioner for finite element discretizations of heterogeneous Stokes flow, Computer Methods in Applied Mechanics and Engineering, 290, 496–523, https://doi.org/10.1016/j.cma.2015.03.014, 2015.

Moresi, L., Quenette, S., Lemiale, V., Mériaux, C., Appelbe, B., and Mühlhaus, H.-B.: Computational approaches to studying non-linear dynamics of the crust and mantle, Physics of the Earth and Planetary Interiors, 163, 69–82, https://doi.org/10.1016/j.pepi.2007.06.009, 2007.

Ratcliff, J. T., Schubert, G., and Zebib, A.: Steady tetrahedral and cubic patterns of spherical shell convection with temperature-dependent viscosity, Journal of Geophysical Research: Solid Earth, 101, 25 473–25 484, https://doi.org/https://doi.org/10.1029/96JB02097, 1996.

Rathgeber, F., Ham, D. A., Mitchell, L., Lange, M., Luporini, F., McRae, A. T. T., Bercea, G.-T., Markall, G. R., and Kelly, P. H. J.: Firedrake: automating the finite element method by composing abstractions, ACM Transactions on Mathematical Software, 43, 1–27, https://doi.org/10.1145/2998441, 2016.

Richards, M. A., Yang, W.-S., Baumgardner, J. R., and Bunge, H.-P.: Role of a low-viscosity zone in stabilizing plate tectonics: Implications for comparative terrestrial planetology, Geochemistry, Geophysics, Geosystems, 2, https://doi.org/https://doi.org/10.1029/2000GC000115, 2001.

https://doi.org/10.5194/egusphere-2026-5839
Preprint. Discussion started: 26 October 2026
© Author(s) 2026. CC BY 4.0 License.

Rudi, J., Malossi, A. C. I., Isaac, T., Stadler, G., Gurnis, M., Staar, P. W. J., Ineichen, Y., Bekas, C., Curioni, A., and Ghattas, O.: An extreme-scale implicit solver for complex PDEs: highly heterogeneous flow in Earth's mantle, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC15), pp. 5:1–5:12, https://doi.org/10.1145/2807591.2807675, 2015.

Rudi, J., Stadler, G., and Ghattas, O.: Weighted BFBT preconditioner for Stokes flow problems with highly heterogeneous viscosity, SIAM Journal on Scientific Computing, 39, S272–S297, https://doi.org/10.1137/16M108450X, 2017.

Saad, Y.: A Flexible Inner-Outer Preconditioned GMRES Algorithm, SIAM Journal on Scientific Computing, 14, 461–469, https://doi.org/10.1137/0914028, 1993.

Schubert, G., Turcotte, D. L., and Olson, P.: Mantle Convection in the Earth and Planets, Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511612879, 2001.

Schuberth, B. S. A., Bunge, H.-P., Steinle-Neumann, G., Moder, C., and Oeser, J.: Thermal versus elastic heterogeneity in high-resolution mantle circulation models with pyrolite composition: High plume excess temperatures in the lowermost mantle, Geochem. Geophys. Geosyst., 10, Q01W01, https://doi.org/10.1029/2008GC002235, 2009.

Stotz, I. L., Iaffaldano, G., and Davies, D. R.: Pressure-driven Poiseuille flow: A major component of the torque-balance governing Pacific plate motion, Geophysical Research Letters, 45, 117–125, https://doi.org/10.1002/2017GL075697, 2018.

Tackley, P. J.: Modelling compressible mantle convection with large viscosity contrasts in a three-dimensional spherical shell using the yin-yang grid, Physics of the Earth and Planetary Interiors, 171, 7–18, https://doi.org/10.1016/j.pepi.2008.08.005, 2008.

Zheng, L., Zhang, H., Gerya, T., Knepley, M., Yuen, D. A., and Shi, Y.: Implementation of a multigrid solver on a GPU for Stokes equations with strongly variable viscosity based on Matlab and CUDA, International Journal of High Performance Computing Applications, 28, 50–60, https://doi.org/10.1177/1094342013478640, 2014.

Zhong, S., Zuber, M. T., Moresi, L., and Gurnis, M.: Role of temperature-dependent viscosity and surface plates in spherical-shell models of mantle convection, Journal of Geophysical Research: Solid Earth, 105, 11 063–11 082, https://doi.org/10.1029/2000JB900003, 2000.

Zhong, S., McNamara, A., Tan, E., Moresi, L., and Gurnis, M.: A benchmark study on mantle convection in a 3-D spherical shell using CitcomS, Geochemistry, Geophysics, Geosystems, 9, Q10 017, https://doi.org/10.1029/2008GC002048, 2008.