SIAM J. SCI. COMPUT.
Vol. 37, No. 5, pp. A2270–A2294

THREE-FIELD BLOCK PRECONDITIONERS FOR MODELS OF COUPLED MAGMA/MANTLE DYNAMICS*

SANDER RHEBERGEN†, GARTH N. WELLS‡, ANDREW J. WATHEN§, AND RICHARD F. KATZ¶

Abstract. For a prescribed porosity, the coupled magma/mantle flow equations can be formulated as a two-field system of equations with velocity and pressure as unknowns. Previous work has shown that while optimal preconditioners for the two-field formulation can be obtained, the construction of preconditioners that are uniform with respect to model parameters is difficult. This limits the applicability of two-field preconditioners in certain regimes of practical interest. We address this issue by reformulating the governing equations as a three-field problem, which removes a term that was problematic in the two-field formulation in favor of an additional equation for a pressure-like field. For the three-field problem, we develop and analyze new preconditioners and we show numerically that they are optimal in terms of problem size and less sensitive to model parameters, compared to the two-field preconditioner. This extends the applicability of optimal preconditioners for coupled mantle/magma dynamics into parameter regimes of physical interest.

Key words. magma dynamics, mantle dynamics, finite element method, preconditioners

AMS subject classifications. 65F08, 76M10, 86A17, 86-08

DOI. 10.1137/14099718X

1. Introduction.

In this paper we consider numerical methods to efficiently solve the linear system arising from the discretization of the equations for coupled magma/mantle dynamics. These partial differential equations, derived by McKenzie [12], model the two-phase flow of partially molten regions of the Earth's mantle. High ambient temperatures enable slowly creeping flow of crystalline mantle rock and also permit melting of certain mantle minerals. Melting produces magma that resides within an interconnected network of pores amid the mantle grains. The governing equations describe the creeping flow of the high-viscosity, solid mantle matrix and the porous flow of the low-viscosity magma. Although both the magma and the mantle are individually incompressible, the two-phase mixture permits *compaction*: nonzero convergence of the solid flux is balanced by nonzero divergence of the magma flux (or vice versa). Compaction therefore expels (or imbibes) magma locally, changing the volume fraction of magma, termed the porosity. Compaction flow is associated with a bulk viscosity and compaction stresses; it gives rise to many of the interesting features of the coupled dynamics. In a typical strategy for computing these dynamics,

*Submitted to the journal's Methods and Algorithms for Scientific Computing section November 24, 2014; accepted for publication (in revised form) June 8, 2015; published electronically September 10, 2015. This work was supported by grants NE/I026995/1 and NE/I023929/1 from the UK Natural Environment Research Council.

http://www.siam.org/journals/sisc/37-5/99718.html

†Department of Applied Mathematics, University of Waterloo, 200 University Ave. W, Waterloo, Ontario, Canada N2L 3G1 (srheberg@uwaterloo.ca).

‡Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom (gnw20@cam.ac.uk).

§Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, United Kingdom (andy.wathen@maths.ox.ac.uk).

¶Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, United Kingdom (richard.katz@earth.ox.ac.uk).

solutions for the solid velocity field and the magma pressure field are obtained for a fixed porosity field; the velocity and pressure are then used to update the porosity. The magma velocity field can be obtained diagnostically from the pressure and solid velocity fields.

Discretization of the elliptic equations for solid velocity and pressure results in a linear system of algebraic equations that can be expressed in a 2×2 block matrix format. Preconditioners are crucial to efficiently solve the resulting system by iterative methods; Rhebergen et al. [14] developed a diagonal block preconditioner for this system and proved optimality with respect to problem size. From numerical experiments, however, it was found that performance of the preconditioner deteriorates at high values of the bulk-to-shear-viscosity ratio. This parameter regime, which corresponds to low values of porosity, is common in coupled magma/mantle dynamics simulations: it is found, for example, at the boundary between unmolten and partially molten mantle, where porosity varies continuously from zero through values 1%. Such situations make the preconditioner in [14] of limited practical use. At low values of porosity, the compaction stresses can become dominant over the shear stresses. In this case, the contribution of a "grad-div" term in the momentum balance equation becomes significant; such terms are known to be problematic for standard multigrid methods. The two-field preconditioner in [14] relied on multigrid methods for the matrix blocks. The manifestation of the problem was increasing Krylov solver iteration counts as the bulk-to-shear-viscosity ratio increased.

In this paper, to circumvent the troublesome grad-div term, we introduce a "compaction pressure" field, as was done by Katz et al. [7] and Keller, May, and Kaus [8], and we reformulate the problem as a three-field system. This approach is also used in nearly incompressible elasticity. Discretizing the model leads to a linear system of equations that may be expressed in a 3×3 block matrix format for which we develop and analyze new block preconditioners in this work. By introducing a compaction pressure field, the size of the system is increased compared to the 2×2 block matrix. The relative increase in degrees of freedom is limited, however, as the degrees of freedom for the compaction pressure (like the degrees of freedom of the fluid pressure) are fewer than the degrees of freedom of the solid velocity. Moreover, we will demonstrate through numerical examples that effective preconditioners for the three-field problem compensate for the addition of an extra scalar field to the problem.

The remainder of this paper is structured as follows. In section 2 we present the two- and three-field governing equations. We describe a weak formulation in section 3, develop and analyze a lower block triangular preconditioner in section 4, and then discuss a diagonal block preconditioner in section 5. In section 6 we verify our analysis by two- and three-dimensional numerical simulations. Conclusions are drawn in section 7.

2. Governing equations. On a domain $\Omega \subset \mathbb{R}^d$, where $1 \leq d \leq 3$, for a given porosity field $\phi \in [0, 1]$ the nondimensional two-phase flow equations that describe coupled magma/mantle dynamics are given by

$$-\nabla \cdot (\eta \mathbf{D}\mathbf{u}) + \nabla p = \nabla \left(\left(\zeta - \tfrac{1}{3}\eta \right) \nabla \cdot \mathbf{u} \right) + \phi \mathbf{e}_3, \tag{2.1a}$$

$$\nabla \cdot \mathbf{u} = \nabla \cdot \left(k \left(\nabla p - \mathbf{e}_3 \right) \right), \tag{2.1b}$$

where $\eta > 0$ is the shear viscosity, $\mathbf{u}$ is the matrix velocity, $\mathbf{D}\mathbf{u} = (\nabla \mathbf{u} + (\nabla \mathbf{u})^T)/2$ is the total strain rate, p is the dynamic pressure, $\zeta > 0$ is the bulk viscosity, $k \geq 0$ is the permeability, and $\mathbf{e}_3$ is the unit vector in the direction aligned with gravity

(i.e., $\mathbf{e}_3 = (0,1)$ when $d = 2$ and $\mathbf{e}_3 = (0,0,1)$ when $d = 3$). Throughout this paper we take the porosity ϕ to be a function of $\mathbf{x} \in \Omega$. Constitutive relations are required for the permeability k, shear viscosity η, and bulk viscosity ζ. For now we just mention that k, η, and ζ are usually functions of the porosity ϕ. For more details on the derivation of the two-phase flow equations (2.1) we refer to McKenzie [12]. The nondimensionalization of these equations is presented in Appendix A.

In Rhebergen et al. [14] we studied (2.1) for the restricted case of constant shear viscosity, constant bulk viscosity, and a spatially variable permeability that is independent of porosity. These simplifications lead to the following system of equations:

$$-\nabla \cdot \mathbf{D}\mathbf{u} + \nabla \tilde{p} = \nabla (\alpha \nabla \cdot \mathbf{u}) + \phi \mathbf{e}_3/\eta, \tag{2.2a}$$

$$\nabla \cdot \mathbf{u} = \nabla \cdot \left(\tilde{k} \left(\nabla \tilde{p} - \mathbf{e}_3/\eta \right) \right), \tag{2.2b}$$

where $\alpha = \zeta/\eta - 1/3$, $\tilde{p} = p/\eta$, and $\tilde{k} = \eta k$. In [14] we developed and analyzed a diagonal block preconditioner for a mixed finite element discretization of (2.2). Combined with a Krylov method, the preconditioner developed in [14] resulted in an optimal solver in terms of the problem size but was not uniform with respect to the model parameters. In particular, as α increased the iteration count for the solver to reach a set tolerance increased. This was attributed to the performance of standard multigrid (geometric and algebraic) when the relative contribution of the $\nabla(\nabla \cdot \mathbf{u})$ term becomes significant [14].

In this paper we develop new preconditioners for a reformulated system of equations in which the $\nabla(\nabla \cdot \mathbf{u})$ term does not appear explicitly. To achieve this we return to (2.1) and introduce the auxiliary variable $p_c = -\zeta \nabla \cdot \mathbf{u}$, which allows us to write (2.1) as

$$-\nabla \cdot \left(\eta \left(\mathbf{D}\mathbf{u} - \tfrac{1}{3} \nabla \cdot \mathbf{u} \mathbb{I} \right) \right) + \nabla p + \nabla p_c = \phi \mathbf{e}_3, \tag{2.3a}$$

$$-\nabla \cdot \mathbf{u} + \nabla \cdot k \nabla p = \nabla \cdot k \mathbf{e}_3, \tag{2.3b}$$

$$-\nabla \cdot \mathbf{u} - \zeta^{-1} p_c = 0. \tag{2.3c}$$

The auxiliary variable p_c is also known as the compaction pressure (see [7, 8], for example). Decomposing the boundary of the domain by $\Gamma_D \cup \Gamma_N = \partial\Omega$, where $\Gamma_D \cap \Gamma_N = \emptyset$, and denoting the outward unit normal vector on $\partial\Omega$ by $\mathbf{n}$, we consider the following boundary conditions:

$$\begin{aligned} \mathbf{u} &= \mathbf{g} && \text{on } \Gamma_D, \\ \eta \mathbf{D}\mathbf{u} \cdot \mathbf{n} - \left(\tfrac{1}{3} \eta \nabla \cdot \mathbf{u} + p + p_c \right) \mathbf{n} &= \mathbf{g}_N && \text{on } \Gamma_N, \\ -k \left(\nabla p - \mathbf{e}_3 \right) \cdot \mathbf{n} &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{2.4}$$

where $\mathbf{g} : \Gamma_D \to \mathbb{R}^d$ and $\mathbf{g}_N : \Gamma_N \to \mathbb{R}^d$ are given boundary data. In the case $\partial\Omega = \Gamma_D$, $\mathbf{g}$ is constructed to satisfy the compatibility condition

$$0 = \int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} \, \mathrm{d}s. \tag{2.5}$$

Note that the compatibility condition implies $\int_\Omega \zeta^{-1} p_c \, \mathrm{d}x = 0$ when $\Gamma_D = \partial\Omega$.

3. Discrete formulation. Assume $\Gamma_D = \partial\Omega$ and, without loss of generality, homogeneous boundary conditions on $\mathbf{u}$. Define the function space $L_0^2 := L_0^2(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q\,dx = 0\}$ and let $\mathbf{X}_h \subset \mathbf{H}_0^1$ and $M_h \subset (H^1 \cap L_0^2)$ be finite dimensional spaces. A mixed finite element weak formulation for (2.3) is then given by the following: find $(\mathbf{u}_h, p_h, p_{ch}) \in \mathbf{X}_h \times M_h \times M_h$ such that

$$a(\mathbf{u}_h, \mathbf{v}) + b(p_h, \mathbf{v}) + b(p_{ch}, \mathbf{v}) = \int_\Omega \phi \mathbf{e}_3 \cdot \mathbf{v}\,\mathrm{d}x \qquad \forall \mathbf{v} \in \mathbf{X}_h, \tag{3.1a}$$

$$b(q, \mathbf{u}_h) - c(p_h, q) = -\int_\Omega k\mathbf{e}_3 \cdot \nabla q\,\mathrm{d}x \quad \forall q \in M_h, \tag{3.1b}$$

$$b(\omega, \mathbf{u}_h) - d(p_{ch}, \omega) = 0 \qquad \forall \omega \in M_h, \tag{3.1c}$$

where

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \eta \mathbf{D}\mathbf{u} : \mathbf{D}\mathbf{v}\,\mathrm{d}x - \int_\Omega \tfrac{1}{3}\eta(\nabla \cdot \mathbf{u})(\nabla \cdot \mathbf{v})\,\mathrm{d}x, \tag{3.2a}$$

$$b(p, \mathbf{v}) = -\int_\Omega p \nabla \cdot \mathbf{v}\,\mathrm{d}x, \tag{3.2b}$$

$$c(p, q) = \int_\Omega k \nabla p \cdot \nabla q\,\mathrm{d}x, \tag{3.2c}$$

$$d(p, \omega) = \int_\Omega \zeta^{-1} p\omega\,\mathrm{d}x. \tag{3.2d}$$

We assume the choice of spaces $\mathbf{X}_h$ and M_h satisfy the inf-sup stability condition but postpone the choice of the finite element spaces until section 6.

Let $u \in \mathbb{R}^{n_u}$ be the vector of discrete velocity with respect to the basis for $\mathbf{X}_h$, and let $p \in N^{n_p} = \{q \in \mathbb{R}^{n_p} | q \neq 1\}$ be the vector of the discrete pressure and $p_c \in N^{n_p}$ the vector of discrete compaction pressure, with respect to the basis for M_h. The discrete system (3.1) can then be written in block matrix form as

$$\begin{bmatrix} K_\eta & G^T & G^T \\ G & -C_k & 0 \\ G & 0 & -Q_\zeta \end{bmatrix} \begin{bmatrix} u \\ p \\ p_c \end{bmatrix} = \begin{bmatrix} f \\ g \\ 0 \end{bmatrix}, \tag{3.3}$$

where K_η, G, C_k, and Q_ζ are the matrices obtained from the discretization of the bi-linear forms $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, $c(\cdot,\cdot)$, and $d(\cdot,\cdot)$, respectively. This is the system for which we wish to develop and deploy effective preconditioners.

4. Three-field block preconditioners. We now formulate and analyze block preconditioners for the system in (3.3). To achieve this, we assume that $0 < \eta, \zeta < \infty$, and $0 \leq k < \infty$ are constants, in which case (3.3) can be rewritten as

$$\underbrace{\begin{bmatrix} \eta K & G^T & G^T \\ G & -kC & 0 \\ G & 0 & -\zeta^{-1}Q \end{bmatrix}}_{\mathcal{A}} \begin{bmatrix} u \\ p \\ p_c \end{bmatrix} = \begin{bmatrix} f \\ g \\ 0 \end{bmatrix}. \tag{4.1}$$

This format of the equations will guide us toward the correct scaling of the different blocks in the preconditioners for the case of nonconstant η, ζ, and k.

To simplify the notation in the following, we introduce the shorthand

$$\bar{S} = GK^{-1}G^T. \tag{4.2}$$

We assume that the spaces $\mathbf{X}_h$ and M_h are chosen such that

$$\ker G^T = \{\mathbf{1}\}, \tag{4.3}$$

where $\{\mathbf{1}\}$ represents the arbitrary constant in the pressure. In this case $\bar{S}$ is invertible (since K is positive definite) on the space complementary to $\{\mathbf{1}\}$. Finite element spaces that are stable for Stokes equations satisfy (4.3). Indeed, so-called inf-sup stable approximation spaces for Stokes have this property uniformly in h. We note that k and ζ are strictly positive and bounded, hence the blocks kC and $\zeta^{-1}Q$ are nonzero. However, if k becomes small as ζ becomes large the second and third rows of $\mathcal{A}$ will approach linear dependence. This degeneracy is a modeling shortcoming of the considered equations.

For the proofs in this section, the following lemma will be used.

LEMMA 4.1. *Let M and N be symmetric and positive definite matrices. If $M-N$ is positive definite, then $N^{-1}-M^{-1}$ is positive definite.*

Proof. See Horn and Johnson [6, Corollary 7.7.4]. ∎

4.1. Theoretical lower block triangular preconditioners. We first consider lower block triangular preconditioners of the form

$$\mathcal{P} = \begin{bmatrix} \eta\mathcal{K} & 0 & 0 \\ G & \mathcal{R} & 0 \\ G & \mathcal{T} & \mathcal{S} \end{bmatrix} \tag{4.4}$$

to precondition (4.1). Our objective is to find expressions for $\mathcal{K}$, $\mathcal{R}$, $\mathcal{S}$, and $\mathcal{T}$ such that the spectrum of the generalized eigenvalue problem

$$\begin{bmatrix} \eta K & G^T & G^T \\ G & -kC & 0 \\ G & 0 & -\zeta^{-1}Q \end{bmatrix} \begin{bmatrix} u \\ p \\ p_c \end{bmatrix} = \Phi \begin{bmatrix} \eta\mathcal{K} & 0 & 0 \\ G & \mathcal{R} & 0 \\ G & \mathcal{T} & \mathcal{S} \end{bmatrix} \begin{bmatrix} u \\ p \\ p_c \end{bmatrix} \tag{4.5}$$

is bounded independent of the mesh cell size h. In this case, the iteration count for a Krylov method applied to the preconditioned system

$$\begin{bmatrix} \eta\mathcal{K} & 0 & 0 \\ G & \mathcal{R} & 0 \\ G & \mathcal{T} & \mathcal{S} \end{bmatrix}^{-1} \begin{bmatrix} \eta K & G^T & G^T \\ G & -kC & 0 \\ G & 0 & -\zeta^{-1}Q \end{bmatrix} \begin{bmatrix} u \\ p \\ p_c \end{bmatrix} = \begin{bmatrix} \eta\mathcal{K} & 0 & 0 \\ G & \mathcal{R} & 0 \\ G & \mathcal{T} & \mathcal{S} \end{bmatrix}^{-1} \begin{bmatrix} f \\ g \\ 0 \end{bmatrix} \tag{4.6}$$

is expected to be optimal in terms of problem size.

The following theorem gives the conditions under which the problem in (4.5) admits only two distinct eigenvalues.

THEOREM 4.2. *Let the matrices K, G, C, and Q and positive constants η, ζ, and k be those given in* (4.1). *In* (4.4), *if $\mathcal{K} = K$ and*

$$\mathcal{R} = -\frac{1}{\sigma\eta}GK^{-1}G^T + \frac{1}{\sigma\eta^2}GK^{-1}G^T\left(\frac{1}{\eta}GK^{-1}G^T + \frac{1}{\zeta}Q\right)^{-1}GK^{-1}G^T - \frac{k}{\sigma}C, \tag{4.7}$$

$$\mathcal{S} = -\left(\frac{1}{\eta}GK^{-1}G^T + \zeta^{-1}Q\right), \tag{4.8}$$

and

$$\mathcal{T} = -\frac{1}{\eta} G K^{-1} G^T, \tag{4.9}$$

where σ is a parameter such that $\sigma < 0$ or $\sigma \in (0,1)$, then $\mathcal{R}$ is invertible and the generalized eigenvalue problem (4.5) *has only two distinct eigenvalues, $\Phi_1 = 1$ and $\Phi_2 = \sigma$. Furthermore, the eigenvectors corresponding to the eigenvalue $\Phi_1 = 1$ have the form $\begin{bmatrix} u^T & 0 & 0 \end{bmatrix}^T$.*

Proof. First we prove that $\mathcal{R}$ is invertible. For this, note that

$$q^T(\eta^{-2}\bar{S})^{-1}(\eta^{-1}\bar{S} + \zeta^{-1}Q)\bar{S}^{-1}q > q^T(\eta^{-2}\bar{S})^{-1}\eta^{-1}\bar{S}\bar{S}^{-1}q = q^T(\eta^{-1}\bar{S})^{-1}q, \tag{4.10}$$

since Q is positive definite. By Lemma 4.1 we therefore find that

$$q^T\eta^{-1}\bar{S}q > q^T\eta^{-1}\bar{S}(\eta^{-1}\bar{S} + \zeta^{-1}Q))^{-1}\eta^{-1}\bar{S}q. \tag{4.11}$$

Since kC is positive semidefinite it is easily seen from (4.7) that if $\sigma > 0$, then $\mathcal{R}$ is negative definite, and if $\sigma < 0$, then $\mathcal{R}$ is positive definite. Hence, $\mathcal{R}$ is invertible.

We now continue by proving that (4.5) has only two distinct eigenvalues. Assuming $\Phi = 1$, (4.5) becomes

$$\eta K u + G^T p + G^T p_c = \eta K u, \tag{4.12a}$$

$$Gu - kCp = Gu + \mathcal{R}p, \tag{4.12b}$$

$$Gu - \zeta^{-1}Qp_c = Gu + \mathcal{T}p + \mathcal{S}p_c. \tag{4.12c}$$

From (4.12a) we find $G^T(p + p_c) = 0$, hence $p = -p_c$, provided both pressures have the same constant average. From (4.12b) we find that $(\mathcal{R} + kC)p = 0$. We need to show that $\mathcal{R} + kC = 0$ is nonsingular, in which case $p = 0$. From the definition of $\mathcal{R}$ in (4.7)

$$\mathcal{R} + kC = -\frac{1}{\sigma\eta}\mathcal{G} + \left(1 - \frac{1}{\sigma}\right) kC, \tag{4.13}$$

where

$$\mathcal{G} = GK^{-1}G^T - GK^{-1}G^T \left(GK^{-1}G^T + \frac{\eta}{\zeta}Q\right)^{-1} GK^{-1}G^T. \tag{4.14}$$

We now show that $\mathcal{G}$ is positive definite. Using the shorthand from (4.2) and defining $\bar{Q} = \eta\zeta^{-1}Q$, we need to show that

$$q^T \left(\bar{S} - \bar{S}^T \left(\bar{S} + \bar{Q}\right)^{-1} \bar{S}\right) q > 0 \quad \forall q \in \mathbb{R}^{n_p}, \tag{4.15}$$

or equivalently

$$\tilde{q}^T \bar{S}^{-1} \tilde{q} - \tilde{q}^T \left(\bar{S} + \bar{Q}\right)^{-1} \tilde{q} > 0 \quad \forall \tilde{q} \in \mathbb{R}^{n_p}, \tag{4.16}$$

where $\tilde{q} = \bar{S}q$. By Lemma 4.1, since $\bar{S} + \bar{Q} - \bar{S} = \bar{Q}$ is positive definite (because $\eta\zeta^{-1} > 0$ and Q is positive definite), the inequality in (4.16) holds, hence $\mathcal{G}$ is positive definite. If $\sigma < 0$, $\mathcal{R} + kC$ is positive definite (since C is positive semidefinite), and if $\sigma \in (0,1)$, then $\mathcal{R} + kC$ is negative definite. It then follows that $p = p_c = 0$ and that $\Phi = 1$ is an eigenvalue of (4.5) with eigenvector $\begin{bmatrix} u^T & 0 & 0 \end{bmatrix}^T$.

Next we assume $\Phi \neq 1$. Expanding the generalized eigenvalue problem (4.5),

$$\eta K u + G^T p + G^T p_c = \Phi \eta K u, \tag{4.17a}$$

$$G u - k C p = \Phi G u + \Phi \mathcal{R} p, \tag{4.17b}$$

$$G u - \zeta^{-1} Q p_c = \Phi G u + \Phi \mathcal{T} p + \Phi \mathcal{S} p_c. \tag{4.17c}$$

From (4.17a),

$$G u = \frac{1}{\eta(\Phi - 1)} G K^{-1} G^T (p + p_c). \tag{4.18}$$

Substituting this expression into (4.17c) and using the definitions of $\mathcal{S}$ (4.8) and $\mathcal{T}$ (4.9) we find

$$(\Phi - 1)\left(\frac{1}{\eta} G K^{-1} G^T p + \left(\frac{1}{\eta} G K^{-1} G^T + \zeta^{-1} Q\right) p_c\right) = 0. \tag{4.19}$$

Since $\Phi \neq 1$, it follows that

$$p_c = -\frac{1}{\eta}\left(\frac{1}{\eta} G K^{-1} G^T + \zeta^{-1} Q\right)^{-1} G K^{-1} G^T p. \tag{4.20}$$

Using (4.18) and (4.20) in (4.17b), we find that

$$\mathcal{R} p = \left(-\frac{1}{\Phi \eta} \bar{S} + \frac{1}{\Phi \eta^2} \bar{S}\left(\frac{1}{\eta} \bar{S} + \frac{1}{\zeta} Q\right)^{-1} \bar{S} - \frac{k}{\Phi} C\right) p. \tag{4.21}$$

From the definition of $\mathcal{R}$ in (4.7) we have $\Phi = \sigma$. ∎

While the choices for $\mathcal{K}$ $\mathcal{R}$, $\mathcal{S}$, and $\mathcal{T}$ in Theorem 4.2 lead to the generalized eigenvalue problem in (4.5) having only two distinct eigenvalues, it does not constitute a computationally useful preconditioner. Computing the inverse of $G K^{-1} G^T$ is not feasible for nontrivial problems. For this reason, we consider in the next section a related, practical preconditioner for (4.1) for large-scale computations.

4.2. Practical lower block triangular preconditioners. Guided by the preconditioner developed in the previous section, we proceed to formulate and analyze related preconditioners that are practical for large-scale simulations. Our objective is to bound the eigenvalues of the preconditioned system independently of the cell size and, if possible, independently of the model parameters.

4.2.1. Construction. To construct a computationally feasible preconditioner, we need to find a suitable approximation for the inverse of $G K^{-1} G^T$. For this we make use of the following lemma.

LEMMA 4.3. *The matrix $G K^{-1} G^T$ is spectrally equivalent to Q:*

$$c_g \leq \frac{\langle G K^{-1} G^T q, q\rangle}{\langle Q q, q\rangle} \leq c^g, \tag{4.22}$$

where c_g and c^g are positive constants independent of h.

Proof. See Elman, Silvester, and Wathen [2, Theorem 5.22]. ∎

Lemma 4.3 suggests that we may replace each occurrence of $GK^{-1}G^T$ in the "theoretical" preconditioner by a weighted pressure mass matrix c_iQ in the expressions for $\mathcal{R}$, $\mathcal{S}$, and $\mathcal{T}$ in Theorem 4.2, resulting in

$$R = -\frac{1}{\sigma\eta}\left(c_1 - \frac{c_2 c_4}{c_3 + \eta\zeta^{-1}}\right) Q - \frac{k}{\sigma} C, \tag{4.23}$$

$$S = -\left(\frac{c_5}{\eta} + \zeta^{-1}\right) Q, \tag{4.24}$$

and

$$T = -\frac{c_6}{\eta} Q, \tag{4.25}$$

respectively. Noting that $\eta\zeta^{-1}$ is positive, for an admissible σ (see Theorem 4.2) we choose to replace R in (4.23) by a spectrally equivalent operator

$$R = -\eta^{-1}Q - kC. \tag{4.26}$$

Similarly, we choose to replace S in (4.24) by the spectrally equivalent

$$S = -\left((2\eta)^{-1} + \zeta^{-1}\right) Q \tag{4.27}$$

(the above factor of two is based on computational experience). We choose to set $\mathcal{T} = T = 0$, which simplifies the preconditioner. We will show, in section 4.2.2, that this simplifies the analysis, without giving up bounds on the spectrum of the preconditioned operator.

We now define a preconditioner for (4.1) of the form

$$\mathcal{P}_t = \begin{bmatrix} \eta\mathcal{K} & 0 & 0 \\ G & \mathcal{R} & 0 \\ G & 0 & \mathcal{S} \end{bmatrix}, \tag{4.28}$$

in which the matrices $\mathcal{K}$, $\mathcal{R}$, and $\mathcal{S}$ satisfy

$$c_k \le \frac{\langle Kq, q\rangle}{\langle \mathcal{K}q, q\rangle} \le c^k, \quad c_r \le \frac{\langle Rq, q\rangle}{\langle \mathcal{R}q, q\rangle} \le c^r, \quad c_s \le \frac{\langle Sq, q\rangle}{\langle \mathcal{S}q, q\rangle} \le c^s \tag{4.29}$$

for R in (4.26) and S in (4.27), and where c_i and c^i in the above are positive constants that are independent of h, k, η, and ζ. In section 6 we will consider a preconditioner of the form in (4.28) in which $\mathcal{K} = K$, $\mathcal{R} = R$, and $\mathcal{S} = S$, with the action of the inverse computed exactly via LU decomposition. We will denote this preconditioner by $\mathcal{P}_t^{\mathrm{LU}}$. We introduce (4.29) into the definition of the preconditioner to permit a wider range of possible preconditioners that can be computationally more efficient. For example, to build an efficient and scalable preconditioner, we consider in section 6 an approximation of inverses of K, R, and S by algebraic multigrid (AMG) cycles, in which $\mathcal{K} = K^{\mathrm{AMG}}$, $\mathcal{R} = R^{\mathrm{AMG}}$, and $\mathcal{S} = S^{\mathrm{AMG}}$. We will denote this preconditioner by $\mathcal{P}_t^{\mathrm{AMG}}$. Multigrid approximations of K, R, and S are spectrally equivalent approximations for the matrices in question [2, Lemma 6.12] and hence satisfy (4.29).

4.2.2. Analysis. In proposing a practical preconditioner, we have thus far relied on spectrally equivalent submatrices for guidance. We now prove that the spectrum of the system of interest, preconditioned by (4.28), where $\mathcal{K}$, $\mathcal{R}$, and $\mathcal{S}$ satisfy (4.29), can be bounded independently of h. Klawonn [9] proved eigenvalue bounds for 2×2 block-triangular preconditioners for a class of saddle point problems. We follow a similar approach to Klawonn [9], but generalized for 3×3 block-triangular preconditioners.

In the following we assume that $c_k > 1$ in (4.29), and hence $K - \mathcal{K}$ is positive definite. This is always possible by appropriate scaling even though $\mathcal{K} = K$ would seem to be the simplest choice. We use this assumption in the analysis; the choice $\mathcal{K} = K$ would lead to significant degeneracy. However, no rescaling is necessary in the numerical simulations in section 6. In preparation for the analysis, we introduce some definitions. Let $\mathcal{A}$ be defined by (4.1) and $\mathcal{P}_t$ by (4.28); then

(4.30)
$$\mathcal{P}_t^{-1}\mathcal{A} = \begin{bmatrix} \mathcal{K}^{-1}K & \eta^{-1}\mathcal{K}^{-1}G^T & \eta^{-1}\mathcal{K}^{-1}G^T \\ -\mathcal{R}^{-1}G\mathcal{K}^{-1}(K-\mathcal{K}) & -\mathcal{R}^{-1}(\eta^{-1}\tilde{S}+kC) & -\mathcal{R}^{-1}\eta^{-1}\tilde{S} \\ -\mathcal{S}^{-1}G\mathcal{K}^{-1}(K-\mathcal{K}) & -\mathcal{S}^{-1}\eta^{-1}\tilde{S} & -\mathcal{S}^{-1}(\eta^{-1}\tilde{S}+\zeta^{-1}Q) \end{bmatrix},$$

where we have used the shorthand

$$\tilde{S} = G\mathcal{K}^{-1}G^T. \tag{4.31}$$

Introducing

$$\mathcal{H} = \begin{bmatrix} \eta(K-\mathcal{K}) & 0 & 0 \\ 0 & -\mathcal{R} & 0 \\ 0 & 0 & -\mathcal{S} \end{bmatrix}, \tag{4.32}$$

we note that

$$\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A} = \begin{bmatrix} \eta(K-\mathcal{K})\mathcal{K}^{-1}K & (K-\mathcal{K})\mathcal{K}^{-1}G^T & (K-\mathcal{K})\mathcal{K}^{-1}G^T \\ G\mathcal{K}^{-1}(K-\mathcal{K}) & \eta^{-1}G\mathcal{K}^{-1}G^T + kC & \eta^{-1}G\mathcal{K}^{-1}G^T \\ G\mathcal{K}^{-1}(K-\mathcal{K}) & \eta^{-1}G\mathcal{K}^{-1}G^T & \eta^{-1}G\mathcal{K}^{-1}G^T + \zeta^{-1}Q \end{bmatrix}. \tag{4.33}$$

We also introduce

$$\tilde{\mathcal{H}} = \begin{bmatrix} \eta K & 0 & 0 \\ 0 & \eta^{-1}\bar{S}+kC & 0 \\ 0 & 0 & \eta^{-1}\bar{S}+\zeta^{-1}Q-\eta^{-1}\bar{S}(\eta^{-1}\bar{S}+kC)^{-1}\eta^{-1}\bar{S} \end{bmatrix}. \tag{4.34}$$

We will consider bounds for $\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A}$ with respect to $\mathcal{H}$ (see also Lemma 3.4 of Klawonn [9]). To find these bounds, we first formulate some intermediate results. We use the notation $A \leq B$ to denote that $B - A$ is symmetric positive semidefinite.

LEMMA 4.4. *Decomposing* $\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A}$ *as*

$$\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A} = \mathcal{L}\mathcal{D}\mathcal{L}^T, \tag{4.35}$$

where

$$\mathcal{L} = \begin{bmatrix} I & 0 & 0 \\ \eta^{-1}GK^{-1} & I & 0 \\ \eta^{-1}GK^{-1} & \eta^{-1}GK^{-1}G^T(\eta^{-1}GK^{-1}G^T+kC)^{-1} & I \end{bmatrix} \tag{4.36}$$

and

$$\mathcal{D} = \begin{bmatrix} \eta(K\mathcal{K}^{-1}K - K) & 0 & 0 \\ 0 & \eta^{-1}\bar{S} + kC & 0 \\ 0 & 0 & \eta^{-1}\bar{S} + \zeta^{-1}Q - \eta^{-1}\bar{S}(\eta^{-1}\bar{S} + kC)^{-1}\eta^{-1}\bar{S} \end{bmatrix}, \tag{4.37}$$

there exist positive constants $\hat{C}_0$, $\hat{C}_1$, *independent of* h, k, η, *and* ζ *such that*

$$\hat{C}_0\tilde{\mathcal{H}} \leq \mathcal{D} \leq \hat{C}_1\tilde{\mathcal{H}}. \tag{4.38}$$

Proof. From (4.29) and $c_k > 1$ it immediately follows that $\hat{C}_0 = \min\{(c_k - 1), 1\}$ and $\hat{C}_1 = \max\{(c^k - 1), 1\}$. ∎

LEMMA 4.5. *Assume* η *and* ζ *are positive bounded constants and defining*

$$\beta_1 = \frac{\eta\zeta^{-1}}{1 + \eta\zeta^{-1}}, \tag{4.39}$$

the eigenvalues of $\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^T$ *are bounded by the extreme eigenvalues of* $\tilde{\mathcal{H}}$*:*

$$\frac{1}{\max\left(4, \frac{6}{\beta_1 \min\left(\frac{1}{c^g}, 1\right)}\right)}\tilde{\mathcal{H}} \leq \mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^T \leq 5\max\left(1, \frac{1}{\beta_1 \min\left(\frac{1}{c^g}, 1\right)}\right)\tilde{\mathcal{H}}, \tag{4.40}$$

where c^g *is given by Lemma* 4.3.

Proof. From

$$\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^T = \begin{bmatrix} \eta K & G^T & G^T \\ G & 2\eta^{-1}\bar{S} + kC & 2\eta^{-1}\bar{S} \\ G & 2\eta^{-1}\bar{S} & 2\eta^{-1}\bar{S} + \zeta^{-1}Q \end{bmatrix} \tag{4.41}$$

we obtain

$$\begin{aligned} x^T\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^T x &= \begin{bmatrix} u \\ p \\ p_c \end{bmatrix}^T \mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^T \begin{bmatrix} u \\ p \\ p_c \end{bmatrix} \\ &\leq \eta u^T K u + 2|p^T G u| + 2|p_c^T G u| + 4\eta^{-1}|p_c^T \bar{S} p| \\ &\quad + p^T\left(2\eta^{-1}\bar{S} + kC\right)p + p_c^T\left(2\eta^{-1}\bar{S} + \zeta^{-1}Q\right)p_c. \end{aligned} \tag{4.42}$$

Applying the Cauchy–Schwarz inequality and Young's inequality $ab \leq a^2/2 + b^2/2$, we find

$$\begin{aligned} |p^T G u| &\leq \tfrac{1}{2}\left(p^T\left(\eta^{-1}\bar{S} + kC\right)p + u^T\eta K u\right), \\ |p_c^T G u| &\leq \tfrac{1}{2}\left(p_c^T\left(\eta^{-1}\bar{S} + \zeta^{-1}Q\right)p_c + u^T\eta K u\right), \\ |p_c^T \bar{S} p| &\leq \tfrac{1}{2}\left(p_c^T \bar{S} p_c + p^T \bar{S} p\right), \end{aligned} \tag{4.43}$$

so that combining (4.42) and (4.43) we obtain

$$\begin{aligned} x^T\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^T x &\leq 3\eta u^T K u + 5p^T\left(\eta^{-1}\bar{S} + kC\right)p \\ &\quad + 5p_c^T\left(\eta^{-1}\bar{S} + \zeta^{-1}Q\right)p_c. \end{aligned} \tag{4.44}$$

From the definition of β_1 (4.39) and using Lemmas 4.1 and 4.3, we find

$$\begin{aligned} p_c^T\tilde{\mathcal{H}}_{33}p_c &= p_c^T\left(\eta^{-1}\bar{S}+\zeta^{-1}Q-\eta^{-1}\bar{S}(\eta^{-1}\bar{S}+kC)^{-1}\eta^{-1}\bar{S}\right)p_c \\ &\geq p_c^T\zeta^{-1}Qp_c \geq p_c^T\beta_1\min\left(\frac{1}{c^g},1\right)\left(\eta^{-1}\bar{S}+\zeta^{-1}Q\right)p_c \end{aligned} \tag{4.45}$$

so that

$$5p_c^T\left(\eta^{-1}\bar{S}+\zeta^{-1}Q\right)p_c \leq \frac{5}{\beta_1\min\left(\frac{1}{c^g},1\right)}p_c^T\tilde{\mathcal{H}}_{33}p_c. \tag{4.46}$$

Combining (4.44) and (4.46) we find the upper bound in (4.40):

$$\begin{aligned} x^T\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^Tx &\leq 3u^T\tilde{\mathcal{H}}_{11}u+5p^T\tilde{\mathcal{H}}_{22}p+\frac{5}{\beta_1\min\left(\frac{1}{c^g},1\right)}p_c^T\tilde{\mathcal{H}}_{33}p_c \\ &\leq 5\max\left(1,\frac{1}{\beta_1\min\left(\frac{1}{c^g},1\right)}\right)x^T\tilde{\mathcal{H}}x. \end{aligned} \tag{4.47}$$

To obtain the lower bound in (4.40) we follow Klawonn [9] and consider

$$\frac{x^T\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^Tx}{x^T\tilde{\mathcal{H}}x}=\frac{y^T\tilde{\mathcal{H}}y}{y^T\mathcal{L}^{-1}\tilde{\mathcal{H}}\mathcal{L}^{-T}y}, \tag{4.48}$$

where $y:=\mathcal{L}^Tx$ was used as substitution. Now,

$$\mathcal{L}^{-1}\tilde{\mathcal{H}}\mathcal{L}^{-T}=\begin{bmatrix}\eta K & -G^T & -G^T+\Lambda^T\\ -G & 2\eta^{-1}\bar{S}+kC & \Xi^T\\ -G+\Lambda & \Xi & \Upsilon\end{bmatrix} \tag{4.49}$$

where

$$\begin{aligned} \Lambda &= \eta^{-1}\bar{S}(\eta^{-1}\bar{S}+kC)^{-1}G,\\ \Xi &= -\eta^{-1}\bar{S}(\eta^{-1}\bar{S}+kC)^{-1}\eta^{-1}\bar{S},\\ \Upsilon &= 2\eta^{-1}\bar{S}+\zeta^{-1}Q-2\eta^{-1}\bar{S}\left(\eta^{-1}\bar{S}+kC\right)^{-1}\eta^{-1}\bar{S}\\ &\quad+\eta^{-1}\bar{S}\left(\eta^{-1}\bar{S}+kC\right)^{-1}\eta^{-1}\bar{S}(\eta^{-1}\bar{S}+kC)^{-1}\eta^{-1}\bar{S}. \end{aligned} \tag{4.50}$$

Similar to the case of the upper bound, using the Cauchy–Schwarz inequality, Young's inequality, and Lemma 4.1 it can be shown that

$$\begin{aligned} y^T\mathcal{L}^{-1}\tilde{\mathcal{H}}\mathcal{L}^{-T}y &= \begin{bmatrix}v\\q\\q_c\end{bmatrix}^T\mathcal{L}^{-1}\tilde{\mathcal{H}}\mathcal{L}^{-T}\begin{bmatrix}v\\q\\q_c\end{bmatrix}\\ &\leq 4\eta v^TKv+4q^T\left(\eta^{-1}\bar{S}+kC\right)q+6q_c^T\left(\eta^{-1}\bar{S}+\zeta^{-1}Q\right)q_c. \end{aligned} \tag{4.51}$$

Using (4.45) we note that

$$6q_c^T\left(\eta^{-1}\bar{S}+\zeta^{-1}Q\right)q_c \leq \frac{6}{\beta_1\min\left(\frac{1}{c^g},1\right)}q_c^T\tilde{\mathcal{H}}_{33}q_c. \tag{4.52}$$

Combining (4.51) and (4.52) we obtain

$$\begin{aligned} y^T\mathcal{L}^{-1}\tilde{\mathcal{H}}\mathcal{L}^{-T}y \leq & 4v^T\tilde{\mathcal{H}}_{11}v + 4q^T\tilde{\mathcal{H}}_{22}q + \frac{6}{\beta_1 \min\left(\frac{1}{c^g},1\right)}q_c^T\tilde{\mathcal{H}}_{33}q_c \\ & \leq \max\left(4, \frac{6}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right) y^T\tilde{\mathcal{H}}y. \end{aligned} \tag{4.53}$$

Using (4.48)

$$x^T\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^Tx \geq \frac{1}{\max\left(4, \frac{6}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right)}x^T\tilde{\mathcal{H}}x, \tag{4.54}$$

which is the lower bound in (4.40). ∎

LEMMA 4.6. *Assume η and ζ are positive bounded constants and let β_1 be given by* (4.39). *There exist positive constants $\tilde{C}_0$, $\tilde{C}_1$, independent of h such that*

$$\tilde{C}_0\tilde{\mathcal{H}} \leq \mathcal{H}\mathcal{P}_t^{-1}\mathcal{A} \leq \tilde{C}_1\tilde{\mathcal{H}}, \tag{4.55}$$

where

$$\tilde{C}_0 = \frac{\hat{C}_0}{\max\left(4, \frac{6}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right)}, \quad \tilde{C}_1 = 5\max\left(1, \frac{1}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right)\hat{C}_1, \tag{4.56}$$

and $\hat{C}_0$ and $\hat{C}_1$ are the constants in Lemma 4.4.

Proof. Combine (4.38), (4.40), and (4.35) to find

$$\frac{\hat{C}_0}{\max\left(4, \frac{6}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right)}\tilde{\mathcal{H}} \leq \hat{C}_0\mathcal{L}\tilde{\mathcal{H}}\mathcal{L}^T \leq \mathcal{H}\mathcal{P}_t^{-1}\mathcal{A} \tag{4.57}$$

and

$$\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A} \leq \hat{C}_1\mathcal{L}\tilde{H}\mathcal{L}^T \leq 5\max\left(1, \frac{1}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right)\hat{C}_1\tilde{\mathcal{H}}, \tag{4.58}$$

from which the lemma follows with

$$\tilde{C}_0 = \frac{\hat{C}_0}{\max\left(4, \frac{6}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right)}, \qquad \tilde{C}_1 = 5\max\left(1, \frac{1}{\beta_1 \min\left(\frac{1}{c^g},1\right)}\right)\hat{C}_1. \tag{4.59}$$ ∎

Building on Lemma 4.6 we now formulate bounds in terms of $\mathcal{H}$.

LEMMA 4.7. *There exist positive constants C_0 and C_1, independent of h, such that*

$$C_0\mathcal{H} \leq \mathcal{H}\mathcal{P}_t^{-1}\mathcal{A} \leq C_1\mathcal{H}. \tag{4.60}$$

Proof. Building on Lemma 4.6, we need to show that $\mathcal{H}$ (4.32) is spectrally equivalent to $\tilde{\mathcal{H}}$ (4.34). We do so by showing spectral equivalence of the corresponding

diagonal blocks in each matrix. It is clear that $\mathcal{H}_{11} = \eta(K - \mathcal{K})$ is spectrally equivalent to $\tilde{\mathcal{H}}_{11} = \eta K$ by (4.29). Next, consider $\mathcal{H}_{22} = -\mathcal{R}$ and $\tilde{\mathcal{H}}_{22} = \eta^{-1}\bar{S} + kC$ and find, using Lemma 4.3 and (4.29),

$$\begin{aligned} \tilde{\mathcal{H}}_{22} &= \eta^{-1}\bar{S} + kC \leq -c_r \max(c^g, 1)\mathcal{R} = -c^{\mathcal{H}_2}\mathcal{R}, \\ \tilde{\mathcal{H}}_{22} &= \eta^{-1}\bar{S} + kC \geq -c^r \min(c_g, 1)\mathcal{R} = -c_{\mathcal{H}_{22}}\mathcal{R}, \end{aligned} \tag{4.61}$$

so that $-c_{\mathcal{H}_{22}}\mathcal{R} \leq \tilde{\mathcal{H}}_{22} \leq -c^{\mathcal{H}_{22}}\mathcal{R}$, meaning that $\tilde{\mathcal{H}}_{22}$ is spectrally equivalent to $\mathcal{H}_{22}$. Finally, consider $\mathcal{H}_{33} = -\mathcal{S}$ and $\tilde{\mathcal{H}}_{33} = \eta^{-1}\bar{S} + \zeta^{-1}Q - \eta^{-1}\bar{S}(\eta^{-1}\bar{S} + kC)^{-1}\eta^{-1}\bar{S}$. We note that, using Lemmas 4.1 and 4.3, and (4.29),

$$\begin{aligned} \tilde{\mathcal{H}}_{33} &= \eta^{-1}\bar{S} + \zeta^{-1}Q - \eta^{-1}\bar{S}(\eta^{-1}\bar{S} + kC)^{-1}\eta^{-1}\bar{S}, \\ &\leq -2c_s \max(c^g, 1)\mathcal{S} = -c^{\mathcal{H}_{33}}\mathcal{S}, \\ \tilde{\mathcal{H}}_{33} &= \eta^{-1}\bar{S} + \zeta^{-1}Q - \eta^{-1}\bar{S}(\eta^{-1}\bar{S} + kC)^{-1}\eta^{-1}\bar{S} \\ &\geq \zeta^{-1}Q = \beta_2\left((2\eta)^{-1} + \zeta^{-1}\right)Q \geq -c^s\beta_2\mathcal{S} = -c_{\mathcal{H}_{33}}\mathcal{S}, \end{aligned} \tag{4.62}$$

where

$$\beta_2 = \frac{\eta\zeta^{-1}}{\frac{1}{2} + \eta\zeta^{-1}}, \tag{4.63}$$

which is positive and bounded because η and ζ are positive and bounded. From (4.62) it therefore follows that $-c_{\mathcal{H}_{33}}\mathcal{S} \leq \tilde{\mathcal{H}}_{33} \leq -c^{\mathcal{H}_{33}}\mathcal{S}$, hence $\tilde{\mathcal{H}}_{33}$ is spectrally equivalent to $\mathcal{S}$. Since $\tilde{\mathcal{H}}_{11}$, $\tilde{\mathcal{H}}_{22}$, and $\tilde{\mathcal{H}}_{33}$ are spectrally equivalent to $\eta(K - \mathcal{K})$, $\mathcal{R}$, and $\mathcal{S}$, respectively, $\mathcal{H}$ is spectrally equivalent to $\tilde{\mathcal{H}}$, and the lemma follows. ∎

THEOREM 4.8. *The spectrum of $\mathcal{P}_t^{-1}\mathcal{A}$ is bounded by the positive constants C_0 and C_1 from Lemma 4.7:*

$$\sigma\left(\mathcal{P}_t^{-1}\mathcal{A}\right) \subset [C_0, C_1]. \tag{4.64}$$

Proof. The proof can be found in Klawonn [9, Theorem 3.5] and is provided here for completeness. The constants C_0 and C_1 of Lemma 4.7 provide the lower and upper bounds for the eigenvalues of the generalized eigenvalue problem

$$\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A}x = \lambda\mathcal{H}x. \tag{4.65}$$

Since $\mathcal{H}$ is nonsingular, the eigenvalues of the above problem are the same as the eigenvalues of

$$\mathcal{P}_t^{-1}\mathcal{A}y = \lambda y. \tag{4.66}$$

∎

We have thus far considered a lower block triangular preconditioner $\mathcal{P}_t$ (4.28). An alternative would be to consider an upper block triangular preconditioner

$$\mathcal{P}_{tU} = \begin{bmatrix} \eta\mathcal{K} & G^T & G^T \\ 0 & \mathcal{R} & 0 \\ 0 & 0 & \mathcal{S} \end{bmatrix}. \tag{4.67}$$

It was noted by Klawonn [9] that since

$$\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A} = \mathcal{A}\mathcal{P}_{tU}^{-1}\mathcal{H}, \tag{4.68}$$

the results in this section also apply to the spectrum of $\mathcal{A}\mathcal{P}_{tU}^{-1}$.

Klawonn [9, Corollary 3.6] shows that $\mathcal{H}^{-1}$ defines an inner product on $\mathbb{R}^{n_u+2n_p}$ and that $\mathcal{P}^{-1}\mathcal{A}$ is symmetric positive definite in this inner product. This implies that one might employ the conjugate gradient method in the $\mathcal{H}^{-1}$ inner product. Due to Theorem 4.8, this then results in optimality in h of the preconditioned CG method. We, however, use the ℓ_2 norm since using the $\mathcal{H}^{-1}$ inner product is not computationally practical. A consequence of this choice is that we will require Krylov methods for nonsymmetric matrices. We consider both GMRES and Bi-CGSTAB.

Klawonn [9, Theorem 3.7] provides bounds on the convergence rate of the GMRES method using a norm equivalent to the $\mathcal{H}^{-1}$-norm, while Theorem 3.8 of [9] provides bounds of the convergence rate of the GMRES method in the ℓ_2-norm. The convergence rate of the GMRES method in the ℓ_2-norm, unfortunately, depends on the condition number of the matrix $\mathcal{H}^{1/2}$. This condition number may depend on h, η, ζ, and k. Our numerical simulations in section 6, however, do not show problem-size dependence and show only slight dependence on the parameters η, ζ, and k. Problem-size independence was also observed in [9]. In particular, the higher the bulk-to-shear-viscosity ratio, the more iterations are required to converge to a given tolerance. This is explained in Lemma 4.6 by noting that for small $\eta\zeta^{-1}$, the constant β_1, defined by (4.39), tends to zero. This implies that $\tilde{C}_0$ and $\tilde{C}_1$ tend to, respectively, zero and infinity, leading to an unbounded spectrum of $\mathcal{H}\mathcal{P}_t^{-1}\mathcal{A}$ in the limit of large bulk-to-shear-viscosity ratio. The simulation results seem less dependent on k. Further discussion of the parameters k, η, and ζ is deferred to section 6.

4.2.3. Variable parameter case. Theorem 4.8 states that the spectrum of $\mathcal{P}_t^{-1}\mathcal{A}$, where $\mathcal{P}_t$ is the preconditioner given by (4.28), is bounded above and below by constants independent of h. This indicates that $\mathcal{P}_t$ will be a good preconditioner for the system (4.1). To achieve this result we assumed k, η, and ζ to be constant and that $\mathcal{K}$, $\mathcal{R}$, and $\mathcal{S}$ satisfy (4.29). For nonconstant physical parameters we propose to replace ηK by K_η and set

$$R = -Q_\eta - C_k, \qquad S = -Q_\eta^\zeta, \tag{4.69}$$

which are such that they reduce to (4.26) and (4.27), respectively, when the physical parameters are constant. Here K_η and C_k are the matrices defined in section 3 and Q_η and Q_η^ζ are the matrices obtained from the discretization of the bilinear forms $d_\eta(\cdot,\cdot)$ and $d_\eta^\zeta(\cdot,\cdot)$, respectively, defined by

$$d_\eta(p,\omega) = \int_\Omega \eta^{-1} p\omega \,\mathrm{d}x, \qquad d_\eta^\zeta(p,\omega) = \int_\Omega \left((2\eta)^{-1} + \zeta^{-1}\right) p\omega \,\mathrm{d}x. \tag{4.70}$$

5. Block diagonal preconditioner. The system in (4.1) is symmetric and therefore the use of MINRES would be allowed so long as the preconditioner is symmetric and positive definite. The advantage of this over using GMRES with the lower block triangular preconditioner of section 4.1 is that less memory is required; the advantage over Bi-CGSTAB is that MINRES is guaranteed to converge (subject to the usual floating-point caveats). These advantages motivate us to consider block diagonal preconditioners for (4.1). In working toward a diagonal preconditioner, ignoring the off-diagonal blocks of (4.28) leads to a preconditioner of the form

$$\mathcal{P}_d^\star = \begin{bmatrix} \eta\mathcal{K} & 0 & 0 \\ 0 & \mathcal{R} & 0 \\ 0 & 0 & \mathcal{S} \end{bmatrix}, \tag{5.1}$$

where $\mathcal{K}$, $\mathcal{R}$, and $\mathcal{S}$ are chosen such that they satisfy (4.29). This preconditioner is symmetric but not positive definite. Multiplying $\mathcal{P}_d^\star$ by the block diagonal matrix $\mathcal{J} = \text{bdiag}(I_u, -I_p, -I_p)$, where $I_u \in \mathbb{R}^{n_u \times n_u}$ and $I_p \in \mathbb{R}^{n_p \times n_p}$ are identity matrices, we obtain the symmetric positive definite preconditioner

$$\mathcal{P}_d = \begin{bmatrix} \eta\mathcal{K} & 0 & 0 \\ 0 & -\mathcal{R} & 0 \\ 0 & 0 & -\mathcal{S} \end{bmatrix}, \tag{5.2}$$

which we propose for use with the MINRES method.

As in the previous section, for nonconstant physical parameters η, ζ, and k, we propose to replace ηK by K_η and we let R and S be given by (4.69).

By contrast to the block-triangular preconditioner, we have not been able to prove boundedness of the spectrum for this block diagonal preconditioner. Instead, in section 6 we investigate its properties by computation.

6. Numerical simulations. In this section we examine numerically the performance of the proposed preconditioners. For both the blocktriangular and block diagonal preconditioners we consider two approaches for the action of the inverse of the matrices $\mathcal{K}$, $\mathcal{R}$, and $\mathcal{S}$: LU decomposition (denoted by $\mathcal{P}^{LU}$) and AMG (denoted $\mathcal{P}^{\text{AMG}}$). We use the LU decomposition as the reference preconditioner to which the AMG preconditioner can be compared. The LU-based preconditioner, however, is not suitable for large-scale computations. We also remark that with the LU preconditioner, the constants in (4.29) are all equal to unity. When using the AMG-based preconditioners, we use smoothed aggregation AMG for the $\mathcal{K}$ block and classical AMG for the $\mathcal{R}$ and $\mathcal{S}$ blocks. We use a single multigrid V-cycle, and unless otherwise stated, for smoothed aggregation we use four applications of a Chebyshev smoother with one symmetric Gauss–Seidel iteration for each Chebyshev application. Although we use AMG to compute the action of the inverse of the matrices $\mathcal{K}$, $\mathcal{R}$, and $\mathcal{S}$, we note that any spectrally equivalent operator may be used.

In all test cases we use P^2-P^1-P^1 continuous Lagrange finite elements on simplices; this combination satisfies (4.3) and will be inf–sup stable since $\zeta^{-1} > 0$. More is needed when $\zeta^{-1} = 0$. We terminate the solver once a relative true residual of 10^{-8} is reached. In the case of GMRES, we use a restarted method with restarts after k iterations. We denote this by GMRES(k).

All experiments have been performed using libraries from the FEniCS Project [11, 10] and the block preconditioning support from PETSc [1]. For smoothed aggregation AMG, we use the library ML [3], while classical AMG is used via the BoomerAMG library [5]. The source code for reproducing all examples is freely available in the supporting material [13].

6.1. Constant bulk and shear viscosity test case in two dimensions. In this test case we consider the simplified two-phase flow equations in (2.2). For the parameters in (2.3), we set $\eta = 1$, $\zeta = \alpha + 1/3$, and

$$k = \frac{k^* - k_*}{4\tanh(5)} \Big(\tanh(10x - 5) + \tanh(10z - 5) + \frac{2(k^* - k_*) - 2\tanh(5)(k_* + k^*)}{k_* - k^*} + 2 \Big), \tag{6.1}$$

where k_* and k^* are parameters that control the maximum and minimum values of k is a domain. We ignore the buoyancy terms but prescribe a source term $\mathbf{f}$ in (2.3a). The Dirichlet boundary condition and the source term are constructed such that the exact solution is

$$u_x = k\partial_x p + \sin(\pi x)\sin(2\pi z) + 2, \tag{6.2}$$

$$u_z = k\partial_z p + \frac{1}{2}\cos(\pi x)\cos(2\pi z) + 2, \tag{6.3}$$

$$p = -\cos(4\pi x)\cos(2\pi z). \tag{6.4}$$

We consider this test case on a structured triangular mesh of the unit square $\Omega = [0,1]^2$. This test case was studied in [14].

6.1.1. Iteration counts for different preconditioners. We first consider the case of $(k_*, k^*) = (0.5, 1.5)$ for the diagonal and block triangular preconditioners. Table 1 presents the iteration counts for the two-field preconditioner from [14] and the three-field block diagonal preconditioner from this work, using MINRES. In both cases, LU and AMG versions are considered. The results in Table 1 show that the LU versions of the preconditioners are optimal, with the three-field version showing some sensitivity to the α parameter. Of greater interest is the performance of the AMG-based preconditioners. In this case, the iteration count of the three-field preconditioner is largely insensitive to the problem size or the α parameter. For the two-field AMG preconditioner, the iteration count has a strong dependency on α. It is this observation from [14] that motivated the present work.

Tables 2 and 3 present the number of iterations required to converge for the block triangular preconditioner using Bi-CGSTAB and GMRES(100), respectively. Compared to the three-field diagonal preconditioner, the triangular preconditioner requires fewer iterations. For the AMG-based cases, the three-field preconditioner requires two to four times fewer iterations than the two-field preconditioner when $\alpha = 100$ and $\alpha = 1000$. Noteworthy is that the Bi-CGSTAB tests require fewer iterations than the GMRES(100) tests. Overall, the three-field preconditioners show less sensitivity to the parameter α than the two-field preconditioner of Rhebergen et al. [14] (see Table 1).

6.1.2. Observed convergence rates. To understand the behavior of the two- and three-field preconditioners, it is helpful to examine the change in the residual with iteration count. This is shown in Figure 1 for the AMG-based two- and three-field block diagonal preconditioners with MINRES and the block triangular preconditioner with Bi-CGSTAB and GMRES(100). We observe that the residual with the two-field preconditioner reduces rapidly to a relative residual of approximately 10^{-4}, at which point the convergence slows. This behavior is not observed with the three-field preconditioners.

We note that dropping just three orders in magnitude in the residual is a criterion for convergence that is commonly used. With a relative tolerance of 10^{-3}, the performance of the two-field preconditioner would appear to be very good and we would draw substantially different conclusions on the relative merits of the two- and three-field formulations. However, we have performed tests that show that a much tighter tolerance is required to maintain the convergence rates to the exact solution with mesh refinement. We discuss this in Appendix B.

TABLE 1

Iteration counts for constant bulk and shear viscosity tests using the two- and three-field block diagonal preconditioners for different values of α. The number of cells in the mesh is $2N$.

$\mathcal{P}_{d2}^{LU}$/MINRES

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	8	8	8	7	7	5
64^2	8	8	8	7	7	5
128^2	8	8	7	7	7	5
256^2	7	6	6	6	7	4

$\mathcal{P}_{d3}^{LU}$/MINRES

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	8	15	22	33	39	39
64^2	8	15	21	33	37	39
128^2	8	15	21	33	37	39
256^2	8	13	21	33	39	39

$\mathcal{P}_{d2}^{AMG}$/MINRES

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	19	19	23	40	93	238
64^2	23	25	29	48	115	289
128^2	26	29	33	57	133	338
256^2	30	32	36	66	155	388

$\mathcal{P}_{d3}^{AMG}$/MINRES

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	29	25	35	60	73	82
64^2	34	29	39	66	84	73
128^2	38	32	43	73	97	84
256^2	43	36	46	78	108	95

TABLE 2

Iteration counts for the constant bulk and shear viscosity tests using the three-field block triangular preconditioner and Bi-CGSTAB for different values of α. The number of cells in the mesh is $2N$.

$\mathcal{P}_{t}^{LU}$/Bi-CGSTAB

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	2	5	7	10	12	12
64^2	2	4	7	11	13	13
128^2	2	4	7	11	13	13
256^2	2	4	7	11	13	14

$\mathcal{P}_{t}^{AMG}$/Bi-CGSTAB

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	6	7	10	20	25	27
64^2	8	8	12	22	28	34
128^2	10	9	12	23	36	41
256^2	11	10	12	27	41	50

TABLE 3

Iteration counts for the constant bulk and shear viscosity tests using the three-field block triangular preconditioner and GMRES(100) for different values of α. The number of cells in the mesh is $2N$.

$\mathcal{P}_t^{LU}$/GMRES(100)

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	4	8	12	19	21	21
64^2	4	8	12	19	21	22
128^2	4	8	12	19	22	23
256^2	4	8	12	18	22	23

$\mathcal{P}_t^{AMG}$/GMRES(100)

N	$\alpha = -\frac{1}{3}$	$\alpha = 0$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
32^2	11	13	18	33	41	45
64^2	13	14	21	39	50	54
128^2	15	16	24	42	58	48
256^2	18	18	27	48	66	58

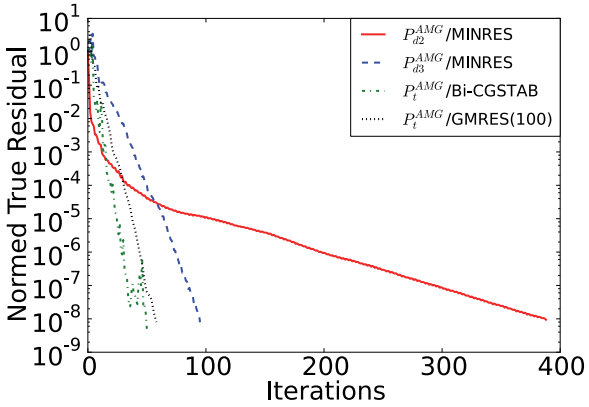


FIG. 1. *Residual decrease using preconditioned MINRES (with $\mathcal{P}_{d2}^{AMG}$ and $\mathcal{P}_{d3}^{AMG}$) and preconditioned Bi-CGSTAB and GMRES(100) (with $\mathcal{P}_t^{AMG}$) for the unit square test of section 6.1. In all cases a mesh size of $N = 256^2$ and $\alpha = 1000$ are used.*

6.2. Variable bulk and shear viscosity test case in two dimensions. In this test case we consider a manufactured solution for the prescribed porosity field

$$\phi = \frac{1}{2}(\phi_* + \phi^*) + \frac{1}{2}(\phi^* - \phi_*)\cos(4\pi(x\sin(\pi/6) + z\cos(\pi/6))),$$

where ϕ_* and ϕ^* are prescribed and $\phi_* \le \phi \le \phi^*$. Let η, ζ, and k be given by

$$k = \frac{R^2}{r_\zeta + 4/3}\left(\frac{\phi}{\phi_0}\right)^m, \quad \eta = 2\exp(-\lambda(\phi - \phi_0)), \quad \zeta = r_\zeta\left(\frac{\phi}{\phi_0}\right)^{-1}, \tag{6.5}$$

where $R = \delta/H$, with δ the reference compaction length $\delta = \sqrt{(r_\zeta + 4/3)\eta_0 k_0/\mu}$, μ is the (constant) melt viscosity, k_0 is the characteristic permeability, η_0 is the characteristic shear viscosity, $r_\zeta = \zeta_0/\eta_0$, with ζ_0 the characteristic bulk viscosity,

TABLE 4

Iteration counts for the block diagonal preconditioners using MINRES for the variable viscosity test in two dimensions. The number of cells in the mesh is $2N$. A dash indicates that we could not compute a converged solution to breakdown of the solver.

	$\mathcal{P}_{d2}^{LU}$/MINRES			$\mathcal{P}_{d2}^{AMG}$/MINRES		
N	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$
32^2	76	-	-	317	-	-
64^2	71	-	-	382	-	-
128^2	70	-	-	428	-	-
256^2	67	-	-	463	-	-

	$\mathcal{P}_{d3}^{LU}$/MINRES			$\mathcal{P}_{d3}^{AMG}$/MINRES		
N	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$
32^2	217	218	219	249	251	251
64^2	223	226	226	227	229	229
128^2	216	222	222	227	243	243
256^2	213	244	247	239	297	299

TABLE 5

Iteration counts for the block triangular preconditioner using Bi-CGSTAB for the variable viscosity test in two dimensions. The number of cells in the mesh is $2N$.

	$\mathcal{P}_{t}^{LU}$/Bi-CGSTAB			$\mathcal{P}_{t}^{AMG}$/Bi-CGSTAB		
N	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$
32^2	72	75	68	71	70	69
64^2	63	69	65	61	60	61
128^2	69	67	68	51	52	51
256^2	69	61	54	48	64	72

and H is a length scale. Furthermore, ϕ_0 is the characteristic porosity and m and λ are constants. We choose $m = 2$, $\lambda = 27$, $r_\zeta = 5/3$, $R = 0.1$, and $\phi_0 = 0.05$. As before, we neglect buoyancy and add a source term $\mathbf{f}$ to the right-hand side of (2.3a). Again the Dirichlet boundary condition and the source term are such that the exact solution for the velocity $\mathbf{u}$ and the pressure p are given by (6.2), (6.3), and (6.4). The approximate solution is computed on a structured triangular mesh of the unit square, $\Omega = [0, 1]^2$. For the tests in this section, we fix $\phi^* = 0.3$ and vary ϕ_*.

To compare more fairly the three-field block diagonal preconditioner $\mathcal{P}_{d3}$ with the two-field block diagonal preconditioner $\mathcal{P}_{d2}$ introduced in [14], we slightly modify the two-field block preconditioner of [14] to include a porosity dependence. This modification is described in Appendix C.

In Table 4 we record the number of iterations for the block diagonal preconditioners with MINRES for different values of ϕ_*. For the AMG-based preconditioner, the iteration count is lower for the three-field preconditioner. For low values of ϕ_* we could not compute converged solutions with the two-field preconditioner. Tables 5 and 6 record the number of iterations for the block triangular preconditioner using Bi-CGSTAB and GMRES(100), respectively. We observe minimal sensitivity to ϕ_* and, again, the iteration count for Bi-CGSTAB is significantly lower than for GMRES(100). For both Bi-CGSTAB and GMRES(100), the iteration counts with the block triangular preconditioner are significantly less than for the two- and three-field block diagonal preconditioners with MINRES.

TABLE 6

Iteration counts for the block triangular preconditioner using GMRES(100) for the variable viscosity test in two dimensions. The number of cells in the mesh is $2N$.

	$\mathcal{P}_t^{LU}$/GMRES(100)			$\mathcal{P}_t^{AMG}$/GMRES(100)		
N	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$	$\phi_* = 10^{-3}$	$\phi_* = 10^{-5}$	$\phi_* = 0$
32^2	124	112	123	121	115	115
64^2	108	118	114	92	94	94
128^2	113	104	98	85	86	86
256^2	96	104	104	81	102	102

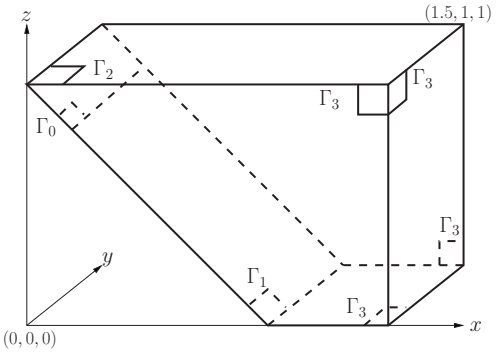


FIG. 2. *Description of a wedge in a three-dimensional subduction zone used for the test cases described in sections* 6.3 *and* 6.4.

6.3. Constant bulk and shear viscosity test case in three dimensions. In this test case we consider a subduction zone setting with a geometry described in Figure 2 (this is the same geometry used in [14]). Boundary subdomains are defined as $\Gamma_1 = \{\mathbf{x} \,|\, x + z = 1\}$, $\Gamma_2 = \{\mathbf{x} \,|\, z = 1\}$, $\Gamma_3 = \partial\Omega \backslash (\Gamma_1 \cup \Gamma_2)$. In this test case we solve the simplified two-phase flow equations in (2.2). For the parameters in (2.3), we set $\eta = 1$, $\alpha = 1000$, $\zeta = \alpha + 1/3$, $k = 0.9(1 + \tanh(-2r))$, with $r = \sqrt{x^2 + z^2}$, and $\phi = 0.01$. This test case was studied also in [14].

We apply the boundary conditions given by (2.4) with $\mathbf{g} = (1/\sqrt{2}, 0.1, -1\sqrt{2})$ and $\mathbf{g} = (0, 0, 0)$ on, respectively, Γ_1 and Γ_2, and $\mathbf{g}_N = (0, 0, 0)$ on Γ_3.

We compute the solution to this test case on three unstructured meshes with an increasing number of degrees of freedom. We only consider the AMG-based preconditioners. In Table 7 we present the number of iterations needed for convergence. The reduced iteration counts for the three-field formulation, relative to the two-field formulation, is clear. This is especially so for the Bi-CGSTAB case.

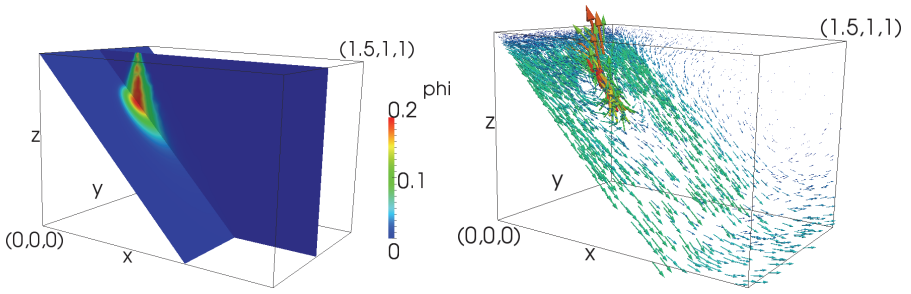
6.4. Porosity-dependent test case in three dimensions. In this test case we again consider the subduction zone-like domain in Figure 2. We set $\Gamma_1 = \{\mathbf{x} \,|\, x + z = 1, x > 0.1\}$ and solve (2.3) with the constitutive relations in (6.5), although we slightly modify the bulk viscosity. For the bulk viscosity we use

$$\zeta_{\text{mod}}^{-1} = \begin{cases} \zeta^{-1} & \text{if } \zeta^{-1} > \zeta_c^{-1}, \\ \zeta_c^{-1} & \text{otherwise,} \end{cases} \tag{6.6}$$

with ζ_c^{-1} being a cut-off inverted bulk viscosity. We choose $\zeta_c^{-1} = 10^{-4}$ which roughly corresponds to $\alpha = 1000$ in (2.2). Other constants are chosen as $m = 2$, $\lambda = 27$, $r_\zeta = 5/3$, $R = 0.1$, and $\phi_0 = 0.05$. A modified bulk viscosity (6.6) is needed to

TABLE 7
Iteration counts for the block preconditioners for the three-dimensional wedge domain with constant viscosity.

DOFs	$\mathcal{P}_{d2}^{AMG}$/MINRES	$\mathcal{P}_{d3}^{AMG}$/MINRES	$\mathcal{P}_{t}^{AMG}$/Bi-CGSTAB	$\mathcal{P}_{t}^{AMG}$/GMRES(100)
419,486	694	366	123	193
1,905,881	772	280	71	115
8,493,971	766	258	74	144



(a) Porosity field. (b) Fluid velocity field.

FIG. 3. *The given porosity field and computed fluid velocity in a three-dimensional subduction zone. Test case of section* 6.4.

prevent the system of equations in (2.3) becoming underdetermined; without the modified bulk viscosity both (2.3b) and (2.3c) reduce to $\nabla \cdot \mathbf{u} = 0$ in the limit $\phi \to 0$.

We prescribe the porosity field

$$\phi = (\phi^* - \phi_*) \exp\left(-\frac{(x-x_c)^2 + (y-y_c)^2}{2\omega^2}\right) + \phi_*, \quad \omega = (\omega^* - \omega_*)\frac{z-1}{z_c - 1} + \omega_*, \tag{6.7}$$

with $\phi^* = 0.2$, $\phi_* = 0$, $x_c = 0.3$, $y_c = 0.5$, $z_c = 1 - x_c$, $\omega^* = 0.07$, and $\omega_* = 0.01$. The porosity field is visualized in Figure 3(a). We apply the boundary conditions in (2.4) with $\mathbf{g} = (1/\sqrt{2}, 0.1, -1\sqrt{2})$ on Γ_1 and $\mathbf{g} = (0,0,0)$ on Γ_2, and $\mathbf{g}_N = (0,0,0)$ on Γ_0 and Γ_3. Once the solution $(\mathbf{u}, p, p_c)$ to (2.3) has been computed, the magma velocity may be recovered from

$$\mathbf{u}_f = \mathbf{u}_s - \frac{k}{\phi}(\nabla p - \mathbf{e}_3). \tag{6.8}$$

We depict the computed magma velocity in Figure 3(b).

We compute the solution to this test case on three unstructured meshes with differing numbers of degrees of freedom. It was not possible to compute a converged solution for this test case with the two-field preconditioner (due to the low porosity) and so we consider the three-field preconditioners only. We consider only the AMG-based preconditioners.

Table 8 shows the number of iterations required to meet the convergence tolerance. We observe no pathological growth in the iteration count with mesh refinement; on the contrary, the iteration count tends to drop with mesh refinement. We use unstructured meshes for this test case and speculate that the reduced iteration count for finer meshes could be due to better mesh quality. Noteworthy, again, is the good performance of Bi-CGSTAB.

TABLE 8
Iteration counts for the variable viscosity wedge problem.

DOFs	MINRES	Bi-CGSTAB	GMRES(300)
419,486	1607	320	918
1,905,881	1534	256	609
8,493,971	933	225	769

6.5. Summary of numerical simulations. By numerical simulations we have shown that for high bulk-to-shear-viscosity ratios, the three-field preconditioner is more robust and has superior performance than the two-field preconditioner developed in [14]. This parameter regime, which corresponds to low values of porosity, is common in coupled magma/mantle dynamics simulations. It is exactly in this regime, where compaction stresses dominate over shear stresses, that the two-field preconditioner breaks down. The main reason to use a two-field preconditioner would be because the global system is smaller than when using the three-field preconditioner. However, for practical applications, the advantages of introducing the compaction pressure in the three-field formulation as a new unknown certainly outweigh the increase in size of the global system.

7. Conclusions. We have proposed, analyzed, and numerically tested new preconditioners for a three-field formulation of the flow equations for coupled magma/mantle dynamics. The system of equations can be formulated as a two-field problem, but it was shown numerically in our past work that a diagonal block preconditioner using algebraic multigrid was not uniform with respect to a parameter that modulates compaction stresses. This motivated the development of preconditioners for a three-field version of the equations, in which an extra pressure variable is introduced. Our analysis shows that for a lower block triangular preconditioner, the eigenvalues of the preconditioned operator are independent of the problem size and have a mild sensitivity to the model parameters. The latter issue is associated with a degeneracy of the model as porosity approaches zero. Numerical experiments indicate that the iteration count for solution of the three-field problem with the preconditioner developed here does not grow with problem size and that the sensitivity to model parameters is small. We therefore expect the preconditioners we have presented to be effective for large-scale simulation of realistic subduction zones with large variations in parameters.

Appendix A. Nondimensionalization of the two-phase flow equations. The two-phase flow equations that describe coupled magma/mantle dynamics, as derived by McKenzie [12], are based on mass and momentum conservation. Mass conservation for the solid (matrix) phase and fluid (melt) phase read

$$\partial_t \phi - \nabla \cdot \big((1-\phi)\mathbf{u}_s\big) = 0, \tag{A.1a}$$

$$\nabla \cdot (\mathbf{u}_s + \mathbf{q}) = 0, \tag{A.1b}$$

where ϕ is the porosity, $\mathbf{u}_s$ is the velocity of the solid phase, and $\mathbf{q} = \phi\left(\mathbf{u}_f - \mathbf{u}_s\right)$, where $\mathbf{u}_f$ is the velocity of the melt phase. We have assumed that there is no melting/freezing and we have taken the density of the solid and melt phases to be constant and uniform. Momentum conservation of the melt phase (Darcy's law) reads

$$\mathbf{q} = -\frac{k}{\mu}\nabla\left(p_f + \rho_f g z\right), \tag{A.2}$$

where k is the permeability, μ is the fluid viscosity, p_f is the pressure in the fluid phase, ρ_f is the mass density of the fluid, and g is the constant acceleration due to gravity. Momentum balance for the two-phase mixture reads

$$-\nabla \cdot (2\eta \mathbf{D}\mathbf{u}_s) + \nabla p_f = \nabla \left(\left(\zeta - \frac{2}{3}\eta \right) \nabla \cdot \mathbf{u}_s \right) - \bar{\rho} g \mathbf{e}_3, \tag{A.3}$$

where η is the shear viscosity of the solid, $\mathbf{D}\mathbf{u}_s = \frac{1}{2}\left(\nabla \mathbf{u}_s + (\nabla \mathbf{u}_s)^T\right)$ is the strain rate, ζ is the bulk viscosity, $\mathbf{e}_3$ is the unit vector in the z-direction, $\bar{\rho} = \rho_f \phi + \rho_s(1-\phi)$ is the bulk density, and ρ_s is the matrix mass density.

In this paper we are concerned with the efficient solution of (A.1b), (A.2), and (A.3) for a given porosity field, hence we can discard (A.1a). Decomposing the melt pressure as $p_f = p - \rho_s g z$, where p is the dynamic pressure and $\rho_s g z$ the "lithostatic" pressure, and substituting (A.2) into (A.1b) to eliminate the Darcy flux $\mathbf{q}$, we obtain

$$-\nabla \cdot (2\eta \mathbf{D}\mathbf{u}_s) + \nabla p = \nabla \left(\left(\zeta - \frac{2}{3}\eta \right) \nabla \cdot \mathbf{u}_s \right) + g \Delta\rho \phi \mathbf{e}_3, \tag{A.4a}$$

$$\nabla \cdot \mathbf{u}_s = \nabla \cdot \left(\frac{k}{\mu} \nabla \left(p - \Delta\rho g z \right) \right), \tag{A.4b}$$

where $\Delta\rho = \rho_s - \rho_f$. Constitutive relations are required for the permeability and the shear and bulk viscosities. For now we define

$$k = k_0 k', \quad \eta = \eta_0 \eta', \quad \zeta = \zeta_0 \zeta', \tag{A.5}$$

where k_0, η_0, and ζ_0 are the characteristic permeability, shear viscosity, and bulk viscosity, respectively, and k', η', and ζ' are nondimensional functions that depend on the porosity ϕ.

We nondimensionalize (A.4) using

$$\mathbf{u}_s = u_0 \mathbf{u}_s', \ \mathbf{x} = H\mathbf{x}', \ (\eta, \zeta) = \eta_0(\eta', r_\zeta \zeta'), \ k = k_0 k', \ p = \Delta\rho g H p', \tag{A.6}$$

where primed variables are nondimensional, $r_\zeta = \zeta_0/\eta_0$, u_0 is the velocity scaling given by $u_0 = \Delta\rho g H^2/\eta_0$, and H is a length scale. Dropping the prime notation, the two-phase flow equations (A.4) in nondimensional form are given by

$$-\nabla \cdot (2\eta \mathbf{D}\mathbf{u}_s) + \nabla p = \nabla \left(\left(r_\zeta \zeta - \frac{2}{3}\eta \right) \nabla \cdot \mathbf{u}_s \right) + \phi \mathbf{e}_3, \tag{A.7a}$$

$$\nabla \cdot \mathbf{u}_s = \nabla \cdot \left(\frac{R^2}{r_\zeta + 4/3} k \left(\nabla p - \mathbf{e}_3 \right) \right), \tag{A.7b}$$

where $R = \delta/H$ and $\delta = \sqrt{(r_\zeta + 4/3)\eta_0 k_0/\mu}$ is the reference compaction length [15].

To simplify the notation, we redefine the nondimensional permeability and shear and bulk viscosities as

$$k \leftarrow \frac{R^2}{r_\zeta + 4/3} k, \quad \eta \leftarrow 2\eta, \quad \zeta \leftarrow r_\zeta \zeta, \tag{A.8}$$

TABLE 9
Convergence of the finite element solution for P^2-P^1-P^1 elements on a unit square test with $\alpha = 1$ and $(k_*, k^*) = (0.5, 1.5)$ with different stopping criteria for the linear solver. Here (u, v) are the two velocity components and p is the fluid pressure.

N	$\|u - u_h\|_2$	Rate	$\|v - v_h\|_2$	Rate	$\|p - p_h\|_2$	Rate
			Relative preconditioned residual: 10^{-10}			
16^2	3.48E-02	4.0	2.00E-02	4.2	4.80E-02	1.8
32^2	3.70E-03	3.2	1.95E-03	3.4	1.25E-02	1.9
64^2	4.56E-04	3.0	2.36E-04	3.0	3.16E-03	2.0
128^2	6.00E-05	2.9	3.16E-05	2.9	7.92E-04	2.0
256^2	8.96E-06	2.7	4.95E-06	2.7	1.98E-04	2.0
			Relative preconditioned residual: 10^{-5}			
16^2	3.53E-02	4.0	2.04E-02	4.2	4.91E-02	1.8
32^2	5.75E-03	2.6	4.38E-03	2.2	1.73E-02	1.5
64^2	4.10E-03	0.5	4.00E-03	0.1	1.25E-02	0.5
128^2	4.01E-03	0.0	4.03E-03	-0.0	1.22E-02	0.0
256^2	4.01E-03	0.0	4.05E-03	-0.0	1.23E-02	-0.0

so that (A.7) becomes

$$-\nabla \cdot (\eta \mathbf{D}\mathbf{u}_s) + \nabla p = \nabla \left(\left(\zeta - \frac{1}{3}\eta \right) \nabla \cdot \mathbf{u}_s \right) + \phi \mathbf{e}_3, \tag{A.9a}$$

$$\nabla \cdot \mathbf{u}_s = \nabla \cdot \left(k \left(\nabla p - \mathbf{e}_3 \right) \right). \tag{A.9b}$$

Appendix B. Rates of convergence of the finite element discretization. In this appendix we consider the rate at which the numerical error in the velocity and fluid pressure fields decreases as a function of the cell size. We use a quadratic polynomial approximation for the velocity and a linear polynomial approximation for the pressures and let h be a measure of the cell size.

Table 9 presents the error in the L^2 norm for the two velocity components, (u, v), and fluid pressure, p, and rates at which the errors reduce with decreasing h. We show the errors and rates of convergence when the solver is terminated at (a) a relative preconditioned residual of 10^{-10} and (b) a relative preconditioned residual of 10^{-5}. When the relative preconditioned residual reaches 10^{-10} we observe that the L^2 errors of the velocity and pressure converge at $\mathcal{O}(h^3)$ and $\mathcal{O}(h^2)$, respectively. However, when we compute the error with a relative preconditioned residual of only 10^{-5}, the error stagnates, with no reduction in the error with mesh refinement. Similar behavior is observed for the two-field formulation.

Appendix C. Two-field preconditioner including viscosity. In this appendix we extend the two-field preconditioner of [14] to include a porosity dependence. The idea is based on that of Grinevich and Olshanskii [4], in which we scale the pressure mass matrix by the viscosity. The two-field preconditioner is given by

$$\mathcal{P}_2 = \begin{bmatrix} \tilde{K}_\eta & 0 \\ 0 & Q_\eta + C_k \end{bmatrix}, \tag{C.1}$$

where Q_η, C_k, and $\tilde{K}_\eta$ are the matrices obtained from, respectively, the discretization of the bilinear forms $d_\eta(\cdot,\cdot)$ in (4.70), $c(\cdot,\cdot)$ in (3.2c), and $\tilde{a}(\cdot,\cdot)$, which is defined as

$$\tilde{a}(\mathbf{u},\mathbf{v}) = \int_\Omega \eta \mathbf{D}\mathbf{u} : \mathbf{D}\mathbf{v}\,\mathrm{d}x + \int_\Omega \left(\zeta - \frac{1}{3}\eta\right)(\nabla\cdot\mathbf{u})(\nabla\cdot\mathbf{v})\,\mathrm{d}x. \tag{C.2}$$

Acknowledgments. S. R. thanks T. Keller, L. Alisic, and J. Rudge for helpful discussions. R. F. K. thanks the Leverhulme Trust for support.

REFERENCES

[1] J. Brown, M. G. Knepley, D. A. May, L. C. McInnes, and B. Smith, *Composable linear solvers for multiphysics*, in Proceedings of the 11th International Symposium on Parallel and Distributed Computing (ISPDC), 2012, pp. 55–62.

[2] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers*, Numer. Math. Sci. Comput., Oxford University Press, Oxford, UK, 2005.

[3] M. W. Gee, C. M. Siefert, J. J. Hu, R. S. Tuminaro, and M. G. Sala, *ML* 5.0 *Smoothed Aggregation Users Guide*, Tech. report SAND2006-2649, Sandia National Laboratories, 2006.

[4] P. P. Grinevich and M. A. Olshanskii, *An iterative method for the Stokes-type problem with variable viscosity*, SIAM J. Sci. Comput., 31 (2009), pp. 3939–3978.

[5] V. E. Henson and U. M. Yang, *BoomerAMG: A parallel algebraic multigrid solver and preconditioner*, Appl. Numer. Math., 41 (2002), pp. 155–177.

[6] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, UK, 2013.

[7] R. F. Katz, M. G. Knepley, B. Smith, M. Spiegelman, and E. T. Coon, *Numerical simulation of geodynamic processes with the portable extensible toolkit for scientific computation*, Phys. Earth Planet. In., 163 (2007), pp. 52–68.

[8] T. Keller, D. A. May, and B. J. P. Kaus, *Numerical modelling of magma dynamics coupled to tectonic deformation of lithosphere and crust*, Geophys. J. Int., 195 (2013), pp. 1406–1442.

[9] A. Klawonn, *Block-triangular preconditioners for saddle point problems with a penalty term*, SIAM J. Sci. Comput., 19 (1998), pp. 172–184.

[10] A. Logg, K.-A. Mardal, and G. N. Wells, eds., *Automated Solution of Differential Equations by the Finite Element Method*, Lect. Notes Comput. Sci. Eng. 84, Springer, New York, 2012.

[11] A. Logg and G. N. Wells, *DOLFIN: Automated finite element computing*, ACM Trans. Math. Software, 37 (2010), pp. 20:1–20:28.

[12] D. McKenzie, *The generation and compaction of partially molten rock*, J. Petrol., 25 (1984), pp. 713–765.

[13] S. Rhebergen and G. N. Wells, *Supporting computer code*, http://www.repository.cam.ac.uk/handle/1810/248270 (2015).

[14] S. Rhebergen, G. N. Wells, R. F. Katz, and A. J. Wathen, *Analysis of block-preconditioners for models of coupled magma/mantle dynamics*, SIAM J. Sci. Comput., 36 (2014), pp. A1960–A1977.

[15] Y. Takei and R. F. Katz, *Consequences of viscous anisotropy in a deforming, two-phase aggregate: Part* 1. *Governing equations and linearised analysis*, J. Fluid Mech., 734 (2013), pp. 424–455.