SIAM J. Sci. Comput.
Vol. 45, No. 4, pp. A1789–A1811

WEIGHTED-NORM PRECONDITIONERS FOR A MULTILAYER TIDE MODEL[*]

COLIN J. COTTER[†], ROBERT C. KIRBY[‡], AND HUNTER MORRIS[‡]

Abstract. We derive a linearized rotating shallow water system modeling tides, which can be discretized by mixed finite elements. Unlike previous models, this model allows for multiple layers stratified by density. Like the single-layer case [R. C. Kirby and T. Kernell, *Comput. Math. Appl.*, 82 (2021), pp. 212–227], a weighted-norm preconditioner gives a (nearly) parameter-robust method for solving the resulting linear system at each time step, but the all-to-all coupling between the layers in the model poses a significant challenge to efficiency. Neglecting the inter-layer coupling gives a preconditioner that degrades rapidly as the number of layers increases. By a careful analysis of the matrix that couples the layers, we derive a robust method that requires solving a reformulated system that only involves coupling between adjacent layers. Numerical results obtained using Firedrake [F. Rathgeber et al., *ACM Trans. Math. Software*, 43 (2016), 24] confirm the theory.

Key words. block preconditioner, finite element method, tide models

MSC codes. 65F08, 65N30

DOI. 10.1137/22M150753X

1. Introduction.

Accurate modeling of tides plays a critical role in computational geosciences. Tide models help geologists to understand sediment transport and coastal flooding, and they help oceanographers to study mechanisms for global circulation [17, 29]. Finite element methods offer theoretically and computationally robust and efficient discretizations of these methods, and are especially attractive in handling irregular coastlines or topography [37]. The literature contains many papers [9, 10, 23, 24, 25, 32] studying mixed finite element pairs for discretization of layers of ocean and atmosphere models. Much of this work relates to dispersion relations and conservation principles, although our work in [11, 12] focuses on semidiscrete energy estimates related to the damping and corresponding error analysis, including a very broad class of possible nonlinear damping models.

This past work has focused on single-layer tide models derived under a linearization of the shallow water approximation. Oceans tend to stratify by density according to depth, however, and more involved models can include multiple layers, each of which have different densities and are coupled together via hydrostatic pressure. A derivation of the fully nonlinear multilayer depth-averaged equations can be found in [4, 27]. Among many interesting features, these equations can lose hyperbolicity in situations approaching Kelvin–Helmholtz instability [26]. A further generalization with the number of layers varying spatially appears in [6]. Here, we consider only a linearized model, suitable for tides rather than more general coastal flows, that does not have this difficulty. We propose a mixed finite element discretization of this

[*] Submitted to the journal's Methods and Algorithms for Scientific Computing section July 5, 2022; accepted for publication (in revised form) March 13, 2023; published electronically July 19, 2023.

https://doi.org/10.1137/22M150753X

Funding: The first author was supported by EPSRC grant EP/R029423/1. The second author was supported by NSF grants 1912653 and 1909176.

[†] Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ England (colin.cotter@imperial.ac.uk).

[‡] Department of Mathematics, Baylor University, Waco, TX 76706 USA (robert_kirby@baylor.edu, h_morris@baylor.edu).

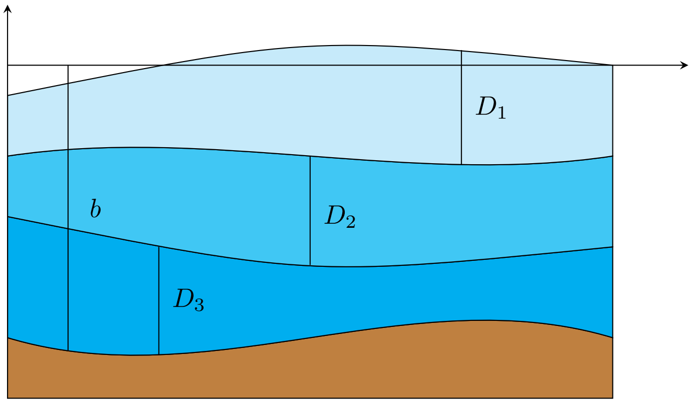

FIG. 2.1. *Example of a one-dimensional fluid with three layers.*

linearized multilayer model and develop effective preconditioners along the lines of those given in the single-layer case in [21]. The all-to-all coupling of the layers presents computational challenges, and special structure of the coupling matrix turns out to be critical. We consider systems of equations arising from *implicit* time stepping rather than the explicit methods in [4, 6, 27]. These methods are better suited for energy conservation in the absence of damping, and can allow large time steps with better stability, but the linear systems required at each time step are challenging to solve.

2. Model and discretization.

We consider a series of layers of fluid inhabiting a domain $\Omega \in \mathbb{R}^2$, with the top layer having thickness D_1, the next layer D_2, and so on until D_N for the bottom layer with bottom boundary $z = b(x, y)$, as shown in Figure 2.1. The density of each layer is denoted by ρ_i, and we assume that $\rho_i < \rho_{i+1}$—the densities strictly increase between layers down a column. Typically, ocean water density varies between 1.02 and 1.07 g/cm^3. Hence, we think of the change in density between the top and bottom layers as being small, and the density difference between two layers as small compared to ρ_1. As a technical assumption that easily covers this case, we posit that

$$\rho_N \le 2\rho_1. \tag{2.1}$$

The horizontal fluid velocity within each layer will be denoted by horizontal velocity $\boldsymbol{u}_i$. We have acceleration due to gravity g, and in this work we take the Coriolis parameter f to be a constant less than 1.

We assume that the pressure is hydrostatic, meaning that the pressure in each layer i satisfies

$$\tfrac{\partial p}{\partial z}|_i = -\rho_i g, \tag{2.2}$$

so $p|_i = -\rho_i g z + c_i$ in each layer.

Using $p = 0$ at the top surface, we have

$$p|_1 = \rho_1 g \left(\sum_{j=1}^{N} D_j + b - z \right). \tag{2.3}$$

Evaluating this at the bottom of the top layer gives

$$p|_1 \left(z = \sum_{j=2}^{N} D_j + b \right) = \rho_1 g D_1. \tag{2.4}$$

Then,

$$p|_2 = \rho_2 g \left(\sum_{j=2}^{N} D_j + b - z + \frac{\rho_1}{\rho_2}\ D_1 \right) = \rho_2 g \left(\sum_{j=1}^{N} D_j + b - z + \frac{\rho_1 - \rho_2}{\rho_2}\ D_1 \right). \tag{2.5}$$

By induction or pattern matching, we have

$$\begin{aligned} p|_i &= \rho_i g \left(\sum_{j=i}^{N} D_j + b - z + \sum_{j=1}^{i-1} \tfrac{\rho_j}{\rho_i} D_j \right) \\ &= \rho_i g \left(\sum_{j=1}^{N} D_j + b - z + \sum_{j=1}^{i-1} \tfrac{\rho_j - \rho_i}{\rho_i} D_j \right). \end{aligned} \tag{2.6}$$

We assume that the motion is layerwise columnar—that is, horizontal velocity is independent of z in each layer. This standard modeling assumption in stratified flow allows columnar flow in one layer to slip past columns in adjacent layers. With this, and dividing by ρ_i, the horizontal component of the momentum equation becomes

$$\begin{aligned} \tfrac{\partial \boldsymbol{u}_i}{\partial t} + \boldsymbol{u_i} \cdot \nabla \boldsymbol{u}_i + f \boldsymbol{u}_i^{\perp} = &- g \nabla \left(\sum_{j=1}^{N} D_j + \sum_{j=1}^{i-1} \frac{\rho_j - \rho_i}{\rho_i} D_j + b \right) \\ &- \frac{C_i(\boldsymbol{u}_i)}{D_i} + \frac{F(t)}{D_i}, \end{aligned} \tag{2.7}$$

where we added a parameterization for bottom drag with $C_i(\boldsymbol{u}_i)$ the damping function, and $F(t)$ is the barotropic tidal forcing. The rationale for the scaling with D_i is that the drag is due to turbulence assumed to occur in the bottom layer only. This turbulent flow exerts an effective damping force proportional to the velocity in the bottom layer, so the depth averaged momentum source is $\int_b^{b+D_N} F(t) dz$, and then we divide by D_i to get the equation for $\boldsymbol{u}$. Mathematically, our framework works with damping in any of the layers, but it is friction in the bottom layer that is the most common setup in practice.

Sometimes a simplified model is used under the rigid lid assumption, in which we assume that $\sum_{j=1}^{N} D_j + b$ is constant. This is relevant because typically $(\rho_j - \rho_i)/\rho_i$ is small, and so there are very fast "barotropic" waves where $\boldsymbol{u}_i$ is independent of i, and much slower "baroclinic" waves where the free surface is more-or-less flat. It is the baroclinic tides that become interesting since they are seen as an "agent for mixing in the deep ocean" [19, 15, 35], which plays a role in setting the large scale ocean circulation [29]. In this work we consider the types of models that are used to investigate the generation of baroclinic tides from the barotropic tide.

Now, we nondimensionalize these equations as follows. We introduce a characteristic vertical length scale H, horizontal length scale L, and velocity scale V. We also introduce a reference density $\overline{\rho}$. Then, we make the change of variables

$$\mathbf{x} = \mathbf{x}' H, \quad t = \frac{H}{V} t', \quad \nabla = \tfrac{1}{H} \nabla', \quad \tfrac{\partial}{\partial t} = \tfrac{V}{H} \tfrac{\partial}{\partial t'}. \tag{2.8}$$

Then, we introduce dimensionless versions of our quantities as

$$C_i = \tfrac{V}{L} C_i', \quad b = H b', \quad D_i = H D_i', \quad \boldsymbol{u}_i = V \boldsymbol{u}_i', \quad \rho_i = \overline{\rho} \rho_i'. \tag{2.9}$$

This gives the following nondimensional equations,

$$\begin{aligned}\tag{2.10} \tfrac{V^2}{L}\left(\tfrac{\partial \boldsymbol{u}_i'}{\partial t'} + \boldsymbol{u}_i'\cdot\nabla'\boldsymbol{u}_i'\right) + Vf\boldsymbol{u_i'}^{\perp} = &-\frac{gH}{L}\nabla'\left(\sum_{j=1}^{N} D_j' + \sum_{j=1}^{i-1}\frac{\rho_j'-\rho_i'}{\rho_i'}D_j' + b'\right)\\ &-\frac{V^2}{LD_i'}\left(C_i'(\boldsymbol{u}_i') - F'(t')\right),\end{aligned}$$

$$\tag{2.11} \tfrac{HV}{L}\tfrac{\partial D_i'}{\partial t'} + \frac{HV}{L}\nabla'\cdot(D_i\boldsymbol{u}_i') = 0,$$

where $F'(t') = \frac{L}{V^2H}F(\frac{L}{V}t')$. Dropping the primes, dividing (2.10) by $\frac{V^2}{L}$, and dividing (2.11) by $\frac{HV}{L}$ produces

$$\begin{aligned}\tag{2.12} \tfrac{\partial \boldsymbol{u}_i}{\partial t} + \boldsymbol{u}_i\cdot\nabla\boldsymbol{u}_i + \epsilon^{-1}\boldsymbol{u}_i^{\perp} = &-Fr^2\nabla\left(\sum_{j=1}^{N} D_j + \sum_{j=1}^{i-1}\frac{\rho_j-\rho_i}{\rho_i}D_j + b\right)\\ &-\frac{1}{D_i}\left(C_i(\boldsymbol{u}_i) - F(t)\right),\end{aligned}$$

$$\tag{2.13} \tfrac{\partial D_i}{\partial t} + \nabla\cdot(D_i\boldsymbol{u}_i) = 0,$$

where $Fr^2 = \frac{gH}{V^2}$ is the square of the Froude number, and $\epsilon^{-1} = \frac{fL}{V}$ is the reciprocal of the Rossby number.

The steady solutions are $\boldsymbol{u}_i = 0$, $i = 1,\ldots,N$, and $D_i = \bar{D}_i =$constant for $i < N$, and $D_N - b = \bar{D}_N - b =$constant. To linearize, we write $D_i = \bar{D}_i + \eta_i$, where $\bar{D}_i$ is the thickness of the layer when the system is at rest. We assume that η_i and $\boldsymbol{u}_i$ are small, retaining only the linear terms in the advection terms as well as replacing D_i by $\bar{D}_i$ in the forcing terms. This gives

$$\tag{2.14} \tfrac{\partial \boldsymbol{u}_i}{\partial t} + \epsilon^{-1}\boldsymbol{u}_i^{\perp} = -Fr^2\nabla\left(\sum_{j=1}^{N}\eta_j + \sum_{j=1}^{i-1}\frac{\rho_j-\rho_i}{\rho_i}\eta_j\right) - \frac{1}{\bar{D}_i}\left(C_i(\boldsymbol{u}_i) - F(t)\right),$$

$$\tag{2.15} \tfrac{\partial \eta_i}{\partial t} + \nabla\cdot\left(\bar{D}_i\boldsymbol{u}_i\right) = 0.$$

Then we make the change of variables $\hat{\boldsymbol{u}}_i = \bar{D}_i\boldsymbol{u}_i$, which makes a kind of momentum rather than velocity of the unknown field. This gives

$$\tag{2.16} \tfrac{1}{\bar{D}_i}\left(\tfrac{\partial \hat{\boldsymbol{u}}_i}{\partial t} + \epsilon^{-1}\hat{\boldsymbol{u}}_i^{\perp}\right) = -Fr^2\nabla\left(\sum_{j=1}^{N}\eta_j + \sum_{j=1}^{i-1}\tfrac{\rho_j-\rho_i}{\rho_i}\eta_j\right) - \tfrac{1}{\bar{D}_i}\left(\hat{C}_i(\hat{\boldsymbol{u}}_i) - F(t)\right),$$

$$\tag{2.17} \tfrac{\partial \eta_i}{\partial t} + \nabla\cdot\hat{\boldsymbol{u}}_i = 0,$$

where $\hat{C}_i(\hat{\boldsymbol{u}}_i) = C_i\left(\frac{\hat{\boldsymbol{u}}_i}{\bar{D}_i}\right)$. Although our model can be formulated with nonlinear damping as in [11], for the rest of the paper we will assume it is linear.

It will be convenient to multiply both sides of (2.16) by ρ_i. Carrying this out, and dropping the circumflexes, gives

$$\tag{2.18} \mu_i\left(\tfrac{\partial \boldsymbol{u}_i}{\partial t} + \epsilon^{-1}\boldsymbol{u}_i^{\perp}\right) = -Fr^2\nabla\left(\sum_{j=1}^{N}\mathcal{A}_{ij}\eta_i\right) - \mu_i\left(C_i(\boldsymbol{u}_i) - F(t)\right),$$

$$\tag{2.19} \tfrac{\partial \eta_i}{\partial t} + \nabla\cdot\boldsymbol{u}_i = 0,$$

where

$$\mathcal{A}_{ij} = \rho_{\min\{i,j\}}. \tag{2.20}$$

For each layer, we let $\mu_i = \frac{\rho_i}{D_i}$.

Let $\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_N \end{bmatrix}$ and $\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_N \end{bmatrix}$. Then we can write (2.18) and (2.19) in matrix-vector notation as follows,

$$\mathcal{M}\tfrac{\partial \boldsymbol{u}}{\partial t} + \epsilon^{-1}\mathcal{M}\boldsymbol{u}^{\perp} + Fr^2\nabla\left(\mathcal{A}\boldsymbol{\eta}\right) + \mathcal{B}\boldsymbol{u} = F, \tag{2.21}$$

$$\tfrac{\partial \boldsymbol{\eta}}{\partial t} + \nabla\cdot\boldsymbol{u} = 0, \tag{2.22}$$

where $\mathcal{M}$ is the diagonal matrix with $\mathcal{M}_{ii} = \mu_i$ and $\mathcal{B}$ is a semidefinite diagonal matrix modeling the damping. Cases of particular interest include the fully definite case, as well as the case where $\mathcal{B}$ vanishes in all except the N, N entry, which corresponds to damping only occurring in the bottom layer.

We let $W = L^2(\Omega)$ be the space of square-integrable functions over Ω with $W_0 = L^2_0(\Omega)$ the subspace of functions with vanishing mean. $V = H(\mathrm{div};\Omega)$ is the space of vector fields over Ω with square-integrable components and divergences. V_0 is the subspace of functions with vanishing normal trace on $\partial\Omega$. We also let $W^N = W \times W \times \cdots \times W$ consist of the N-way Cartesian product of the space with itself, with similar definitions of W_0^N, V^N, and V_0^N. These spaces will be used to represent functions mapping Ω to the disturbances and velocities within each layer.

In the case of vanishing damping, one can apply standard energy techniques similar to wave equations, arriving at stability and well-posedness. With damping present, one has a nonincreasing energy so that we also expect such analysis to carry over. However, our analysis of the single-layer case in [12] gave long-time stability and error analysis for semidiscrete methods by showing the system energy is actually damped exponentially. Similar results should follow readily if the damping is fully positive-definite, but determining the degree to which the results might hold in the semidefinite case is quite interesting.

Throughout, we let $(\cdot,\cdot)$ denote the $(L^2)^N$ inner product with

$$(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega \sum_{i=1}^N \boldsymbol{u}_i\boldsymbol{v}_i dx. \tag{2.23}$$

Without any subscript on the norm, we let $\|\boldsymbol{u}\| = \sqrt{(\boldsymbol{u},\boldsymbol{u})}$ be the standard L^2 norm. For any smooth $\mathcal{S}$ mapping Ω into symmetric and uniformly positive-definite matrices, we also define the $\mathcal{S}$-weighted inner product by

$$(\boldsymbol{u},\boldsymbol{v})_{\mathcal{S}} = (\mathcal{S}\boldsymbol{u},\boldsymbol{v}). \tag{2.24}$$

We assume that $\mathcal{S}$ is uniformly positive-definite over Ω so that

$$\|\boldsymbol{u}\|_{\mathcal{S}} = \sqrt{(\boldsymbol{u},\boldsymbol{u})_{\mathcal{S}}} \tag{2.25}$$

defines a norm equivalent to the L^2 norm with bounds

$$C_{\mathcal{S}}\|\boldsymbol{u}\| \le \|\boldsymbol{u}\|_{\mathcal{S}} \le C^{\mathcal{S}}\|\boldsymbol{u}\| \tag{2.26}$$

for some finite positive constants $C_{\mathcal{S}}$ and $C^{\mathcal{S}}$.

We also assume that $\mathcal{B}$ is bounded in the L^2 norm. That is, there exists some $B^* < \infty$ such that for all $\boldsymbol{u} \in \boldsymbol{V}_h^N$,

$$\|\mathcal{B}\boldsymbol{u}\| \leq B^* \|\boldsymbol{u}\|. \tag{2.27}$$

To arrive at a discrete model, we decompose Ω into a family of quasiuniform meshes $\{\mathcal{T}_h\}_h$ of triangles. For some fixed $k \geq 0$, we let $W_h \subset W$ be the space of all functions whose restrictions to each $T \in \mathcal{T}_h$ are polynomials of degree k, and V_h will consist of a suitable $H(\mathrm{div})$ finite element space, such as the Raviart–Thomas (RT) element [31] or Brezzi–Douglas–Marini (BDM) [8] elements. In the single-layer case, BDM elements may be preferable at small ϵ due to spurious modes appearing with RT [13]. In particular, we assume that the property $\nabla \cdot V_h = W_h$ holds and that there exist suitable commuting projections [5] that would enable stability and error analysis to hold. Decomposition of Ω into quadrilateral meshes is also possible. If the mesh elements are not affine images of a reference square, some accuracy may be lost [2].

We let V_h^N be the finite-dimensional space consisting of vectors of N components, each in V_h, and W_h^N with N components in W_h. By seeking a solution $\boldsymbol{u} : [0,T] \to V_h^N$ and $\boldsymbol{\eta} : [0,T] \to W_h^N$, a Galerkin discretization of (2.21) is

$$\left(\tfrac{\partial \boldsymbol{u}}{\partial t}, \boldsymbol{v}\right)_{\mathcal{M}} + \epsilon^{-1}\left(\boldsymbol{u}^{\perp}, \boldsymbol{v}\right)_{\mathcal{M}} - Fr^2 \left(\boldsymbol{\eta}, \nabla \cdot \boldsymbol{v}\right)_{\mathcal{A}} + \left(\boldsymbol{u}, \boldsymbol{v}\right)_{\mathcal{B}} = \left(F, \boldsymbol{v}\right), \tag{2.28}$$

$$\left(\tfrac{\partial \boldsymbol{\eta}}{\partial t}, \boldsymbol{w}\right) + \left(\nabla \cdot \boldsymbol{u}, \boldsymbol{w}\right) = 0, \tag{2.29}$$

for all $\boldsymbol{v} \in V_h^N$ and $\boldsymbol{w} \in W_h^N$.

To obtain a fully discrete method, we must specify some time-stepping scheme. For example, the implicit midpoint rule is symplectic and, in the damping-free case of $\mathcal{B} = 0$, conserves the system energy exactly for this problem. We assume a constant step size Δt and define discrete time levels $t_n = n\Delta t$. Then, given initial conditions $\boldsymbol{u}_h^0$ and $\boldsymbol{\eta}_h^0$, the solution at each time level is approximated by

$$\begin{aligned}
&\left(\frac{\boldsymbol{u}^{n+1} - \boldsymbol{u}^n}{\Delta t}, \boldsymbol{v}\right)_{\mathcal{M}} + \epsilon^{-1}\left(\left(\boldsymbol{u}^{n+1/2}\right)^{\perp}, \boldsymbol{v}\right)_{\mathcal{M}} \\
&\quad - Fr^2 \left(\boldsymbol{\eta}^{n+1/2}, \nabla \cdot \boldsymbol{v}\right)_{\mathcal{A}} + \left(\mathcal{B}\boldsymbol{u}^{n+1/2}, \boldsymbol{v}\right) = \left(F^{n+1/2}, \boldsymbol{v}\right), \\
&\left(\frac{\boldsymbol{\eta}^{n+1} - \boldsymbol{\eta}^n}{\Delta t}, \boldsymbol{w}\right) + \left(\nabla \cdot \boldsymbol{u}^{n+1/2}, \boldsymbol{w}\right) = 0,
\end{aligned} \tag{2.30}$$

where we define $\boldsymbol{u}_h^{n+1/2} = \frac{1}{2}\left(\boldsymbol{u}_h^n + \boldsymbol{u}_h^{n+1}\right)$ and similarly for $\boldsymbol{\eta}_h^{n+1/2}$. Multiplying through each equation by Δt and moving known data to the right-hand side, we see that a variational problem of the form

$$\begin{aligned}
\left(\boldsymbol{u}, \boldsymbol{v}\right)_{\mathcal{M}} + \epsilon^{-1} k \left(\boldsymbol{u}^{\perp}, \boldsymbol{v}\right)_{\mathcal{M}} - Fr^2 k \left(\boldsymbol{\eta}, \nabla \cdot \boldsymbol{v}\right)_{\mathcal{A}} + k\left(\mathcal{B}\boldsymbol{u}, \boldsymbol{v}\right) &= \left(F_1, \boldsymbol{v}\right), \\
\left(\boldsymbol{\eta}, \boldsymbol{w}\right) + k\left(\nabla \cdot \boldsymbol{u}, \boldsymbol{w}\right) &= \left(F_2, \boldsymbol{w}\right)
\end{aligned} \tag{2.31}$$

must be solved at each time step, where $k > 0$ is some small number related to the time step. This equation is fairly generic—other single-stage methods such as Crank–Nicolson or backward Euler give systems of the same form. A multistage Runge–Kutta method, such as considered in [16] for the wave equation, would give a more complicated system, although the diagonal blocks would have this form.

To simplify the analysis, we define the bilinear form

$$\begin{aligned}
a\left((\boldsymbol{u}, \boldsymbol{\eta}), (\boldsymbol{v}, \boldsymbol{w})\right) = {} & \left(\boldsymbol{u}, \boldsymbol{v}\right)_{\mathcal{M}} + \epsilon^{-1} k \left(\boldsymbol{u}^{\perp}, \boldsymbol{v}\right)_{\mathcal{M}} - Fr^2 k \left(\boldsymbol{\eta}, \nabla \cdot \boldsymbol{v}\right)_{\mathcal{A}} + k\left(\mathcal{B}\boldsymbol{u}, \boldsymbol{v}\right) \\
& + \left(\boldsymbol{\eta}, \boldsymbol{w}\right) + k\left(\nabla \cdot \boldsymbol{u}, \boldsymbol{w}\right),
\end{aligned} \tag{2.32}$$

and the variational problem (2.31) can then be compactly written as finding $(\boldsymbol{u},\boldsymbol{\eta}) \in \boldsymbol{V}_h^N \times \boldsymbol{W}_h^N$ such that

$$a\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right) = (F_1,\boldsymbol{v}) + (F_2,\boldsymbol{w}) \tag{2.33}$$

for all $(\boldsymbol{v},\boldsymbol{w}) \in \boldsymbol{V}_h^N \times \boldsymbol{W}_h^N$.

Now, we cast the discrete variational problem (2.31) into matrix notation. We let $\{\psi_i\}_{i=1}^{\dim V_h}$ be a basis for V_h. Then, if $\boldsymbol{e}^j$ is the canonical basis vector in $\mathbb{R}^n$, 1 in entry j and vanishing in other entries, functions of the form

$$\boldsymbol{\psi}_i^j = \psi_i \boldsymbol{e}^j, \quad 1 \le i \le \dim V_h,\, 1 \le j \le N,$$

form a basis for $\boldsymbol{V}_h^N$. Similarly, we let $\{\phi_i\}_{i=1}^{\dim W_h}$ be a basis for W_h. With

$$\boldsymbol{\phi}_i^j = \phi_i \boldsymbol{e}^j, \quad 1 \le i \le \dim V_h,\, 1 \le j \le N,$$

the set $\{\{\boldsymbol{\phi}_i^j\}_{i=1}^{\dim W_h}\}_{j=1}^N$ forms basis for $\boldsymbol{W}_h$.

In order to define matrices, we impose a total ordering on the basis functions for $\boldsymbol{V}_h^N$ and $\boldsymbol{W}_h^N$. For example, for $1 \le i \le N \dim V_h$, we can find unique i_0, i_1 such that $i = i_0 \dim V_h + i_1$ by integer division/remainder operations and then put

$$\boldsymbol{\Psi}_i = \boldsymbol{\psi}_i^j = \psi_{i_1} \boldsymbol{e}^{i_0}$$

with a similar total ordering for $\{\Phi_i\}_{i=1}^{N \dim V_h}$. This ordering imposes a block structure on the linear system by storing all the degrees of freedom within a layer contiguously.

Before proceeding, we remark on matrix notation, as several different kinds of matrices appear in this paper. Matrices that act across the layers of the tide model, such as $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{M}$ have been denoted in calligraphic letters. Discrete operators on a single layer or, equivalently, those discretizing a bilinear form over V_h and/or W_h will be denoted in italics. To this end, we define:

$$\begin{aligned} M_{ij}^V &= (\psi_j, \psi_i)\,, \\ M_{ij}^{V,\kappa} &= (\kappa\psi_j, \psi_i)\,, \\ \tilde{M}_{ij}^V &= \left(\psi_j^\perp, \psi_i\right), \\ M_{ij}^W &= (\phi_j, \phi_i)\,, \\ D_{ij} &= (\nabla \cdot \psi_j, \phi_i)\,, \\ E_{ij} &= (\nabla \cdot \psi_j, \nabla \cdot \psi_i)\,. \end{aligned} \tag{2.34}$$

Then, we use Roman block lettering to denote discrete operators over $\boldsymbol{V}_h^N$ and/or $\boldsymbol{W}_h^N$. Such needed matrices are:

$$\begin{aligned} \mathrm{M}_{ij}^V &= (\boldsymbol{\Psi}_j, \Psi_i)_{\mathcal{M}}\,, \\ \mathrm{M}_{ij}^W &= (\boldsymbol{\Phi}_j, \boldsymbol{\Phi}_i)\,, \\ \tilde{\mathrm{M}}_{ij} &= \left(\boldsymbol{\Psi}_j^\perp, \boldsymbol{\Psi}_i\right)_{\mathcal{M}}, \\ \mathrm{D}_{ij} &= (\nabla \cdot \boldsymbol{\Psi}_j, \boldsymbol{\Phi}_i)\,, \\ \mathrm{D}_{ij}^{\mathcal{A}} &= (\nabla \cdot \boldsymbol{\Psi}_j, \boldsymbol{\Phi}_i)_{\mathcal{A}}\,, \\ \mathrm{B}_{ij} &= (\mathcal{B}\boldsymbol{\Psi}_j, \boldsymbol{\Psi}_i)\,, \\ \mathrm{E}_{ij} &= (\nabla \cdot \boldsymbol{\Psi}_j, \nabla \cdot \boldsymbol{\Psi}_i)\,, \\ \mathrm{E}_{ij}^{\mathcal{A}} &= (\nabla \cdot \boldsymbol{\Psi}_j, \nabla \cdot \boldsymbol{\Psi}_i)_{\mathcal{A}}\,. \end{aligned} \tag{2.35}$$

Note that the matrices appearing in (2.35) have an important substructure, such as

$$\mathrm{M}^V = diag(M^{V,\mu_1}, \dots, M^{V,\mu_N}). \tag{2.36}$$

The first $N-1$ blocks are constant coefficient and so equal to $\mu_i M^V$. Due to the variable bathymetry, the bottom right block is not, but it is still symmetric and positive-definite. The matrix B is also block diagonal and symmetric semidefinite. If the damping matrix B is full-rank, it is definite. Similarly, M^W, $\tilde{\mathrm{M}}^V$, and D, E are block diagonal. In fact, $\mathrm{W}^W = I \otimes M^W$, $\mathrm{D} = I \otimes D$, and $\mathrm{E} = I \otimes E$, where I is the $N \times N$ identity matrix.

The matrices $\mathrm{D}^{\mathcal{A}}$ and $\mathrm{E}^{\mathcal{A}}$ also have structure, with

$$\begin{aligned} \mathrm{D}^{\mathcal{A}} &= \mathcal{A} \otimes D, \\ \mathrm{E}^{\mathcal{A}} &= \mathcal{A} \otimes E. \end{aligned} \tag{2.37}$$

A Galerkin discretization of (2.31) then gives rise to a block matrix system of the form

$$\begin{bmatrix} \mathrm{M}^V + \epsilon^{-1} k \tilde{\mathrm{M}}^V + k\mathrm{B} & -Fr^2 k \left(\mathrm{D}^{\mathcal{A}}\right)^T \\ k\mathrm{D} & \mathrm{M}^W \end{bmatrix} \begin{bmatrix} \mathrm{u} \\ \eta \end{bmatrix} = \begin{bmatrix} \mathrm{F}_1 \\ \mathrm{F}_2 \end{bmatrix}. \tag{2.38}$$

3. A weighted-norm preconditioner.

Linear systems arising from finite element discretizations are typically solved using iterative methods such as the generalized minimum residual method (hence, GMRES) [34]. These methods have the advantage of requiring only matrix-vector products with the system matrix, but their performance depends strongly on the conditioning of the linear system. The conditioning of the system matrix, and hence number of iterations required for convergence, can degrade as a function of mesh refinement and/or physical parameters. In such cases, it is critical to *precondition* the linear system by premultiplying a linear system

$$Ax = b$$

by some linear operator P^{-1} to obtain the equivalent system

$$P^{-1}Ax = P^{-1}b.$$

One hopes to choose P such that the iterative method converges much faster for $P^{-1}A$ than that of A under the constraint that the cost of applying P^{-1} at each iteration not offset the gains obtained by reducing the iteration count.

When preconditioning finite element linear systems, it can be helpful to choose P as discretizing some simpler differential operator, such as an inner product on the underlying Hilbert space [20, 28]. It is also frequently possible to incorporate physical parameters in the definition of the preconditioner in such a way as to minimize the dependence of the spectral bounds on those parameters. We refer to these as "weighted-norm" preconditioners, and we adopt this perspective here.

In this section, we propose and analyze the matrix

$$\begin{bmatrix} \mathrm{M}^V + Fr^2 k^2 \mathrm{E}^{\mathcal{A}} & 0 \\ 0 & \mathrm{M}^W \end{bmatrix} \tag{3.1}$$

as a preconditioner for (2.38). Because this matrix decouples the momentum and elevation variables, it should be far easier to invert than the original matrix. The M^W block is itself quite simple, just a block diagonal matrix of mass matrices (which can be diagonal in the lowest-order case). However, the top left block couples all of the layer velocities together, and we take a closer look at this block in the following section.

This matrix arises from discretizing the bilinear form

$$b\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right)=\left(\boldsymbol{u},\boldsymbol{v}\right)_{\mathcal{M}}+Fr^2k^2\left(\nabla\cdot\boldsymbol{u},\nabla\cdot\boldsymbol{v}\right)_{\mathcal{A}}+\left(\boldsymbol{\eta},\boldsymbol{w}\right) \tag{3.2}$$

over $\boldsymbol{V}_h^N\times\boldsymbol{W}_h^N$. This bilinear form is equivalent to the standard $H(\mathrm{div})\times L^2$ inner product, with constants dependent upon the physical parameters. We will prove norm equivalence by giving continuity and inf-sup bounds of the bilinear form a in (2.32) with respect to the norm defined by the inner product b.

We first note that the matrix $\mathrm{D}^{\mathcal{A}}$ appears in the first row of the system matrix, but D in the second. Also, the two blocks are scaled differently with respect to the Froude number. This structural asymmetry, complicates the analysis. Rather than scaling the actual system to be solved, we can give an analysis for an equivalent pair of bilinear forms. To motivate this alternate pair, we rewrite the preconditioned matrix,

$$\begin{aligned}&\begin{bmatrix}\mathrm{M}^V+Fr^2k^2\mathrm{E}^{\mathcal{A}} & 0\\ 0 & \mathrm{M}^W\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{M}^V+\epsilon^{-1}k\tilde{\mathrm{M}}^V+kB & -Fr^2k\left(\mathrm{D}^{\mathcal{A}}\right)^T\\ k\mathrm{D} & \mathrm{M}^W\end{bmatrix}\\ &=\begin{bmatrix}\mathrm{M}^V+Fr^2k^2\mathrm{E}^{\mathcal{A}} & 0\\ 0 & Fr^2\mathrm{M}^{W,\mathcal{A}}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{M}^V+\epsilon^{-1}k\tilde{\mathrm{M}}^V+kB & -Fr^2k\left(\mathrm{D}^{\mathcal{A}}\right)^T\\ Fr^2k\mathrm{D}^{\mathcal{A}} & Fr^2\mathrm{M}^{W,\mathcal{A}}\end{bmatrix},\end{aligned} \tag{3.3}$$

where we have inserted the identity, written as

$$\begin{bmatrix}I & 0\\ 0 & Fr^2\mathcal{A}\otimes I\end{bmatrix}^{-1}\begin{bmatrix}I & 0\\ 0 & Fr^2\mathcal{A}\otimes I\end{bmatrix}$$

between the two matrices on the left-hand side.

The second matrix on the right-hand side discretizes the bilinear form

$$\begin{aligned}\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right)=\ &\left(\boldsymbol{u},\boldsymbol{v}\right)_{\mathcal{M}}+\epsilon^{-1}k\left(\boldsymbol{u}^{\perp},\boldsymbol{v}\right)_{\mathcal{M}}-Fr^2k\left(\boldsymbol{\eta},\nabla\cdot\boldsymbol{v}\right)_{\mathcal{A}}+k\left(\mathcal{B}\boldsymbol{u},\boldsymbol{v}\right)\\ &+Fr^2\left(\boldsymbol{\eta},\boldsymbol{w}\right)_{\mathcal{A}}+Fr^2k\left(\nabla\cdot\boldsymbol{u},\boldsymbol{w}\right)_{\mathcal{A}},\end{aligned} \tag{3.4}$$

while the first discretizes the weighted inner product

$$\hat{b}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right)=\left(\boldsymbol{u},\boldsymbol{v}\right)_{\mathcal{M}}+Fr^2k^2\left(\nabla\cdot\boldsymbol{u},\nabla\cdot\boldsymbol{v}\right)_{\mathcal{A}}+Fr^2\left(\boldsymbol{\eta},\boldsymbol{w}\right)_{\mathcal{A}}. \tag{3.5}$$

We further define the $\|\cdot\|_{\hat{b}}$ norm on $\boldsymbol{V}_h^N\times\boldsymbol{W}_h^N$ by

$$\|(\boldsymbol{u},\boldsymbol{\eta})\|_{\hat{b}}=\sqrt{\hat{b}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{u},\boldsymbol{\eta})\right)}. \tag{3.6}$$

Because of equality (3.3), GMRES iteration for the matrix associated with (2.32) preconditioned by that from (3.2) is exactly equivalent to that obtained from the matrices for (3.4) and (3.5). We demonstrate norm equivalence for the latter pair.

THEOREM 3.1. *For all* $(\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})$ *in* $\boldsymbol{V}_h^N\times\boldsymbol{W}_h^N$,

$$\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right)\leq C\left\|(\boldsymbol{u},\boldsymbol{\eta})\right\|_{\hat{b}}\left\|(\boldsymbol{v},\boldsymbol{w})\right\|_{\hat{b}}, \tag{3.7}$$

where

$$C=\max\left\{2,1+\tfrac{k}{\epsilon}+\frac{kB^*}{C_{\mathcal{M}}^2}\right\}. \tag{3.8}$$

Proof. Let $(\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w}) \in \boldsymbol{V}_h^N \times \boldsymbol{W}_h^N$ be given. Then, applying the Cauchy–Schwarz inequality and noting $\cdot^\perp$ is pointwise an isometry, we have

$$\begin{aligned}\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right) =& \left(\boldsymbol{u},\boldsymbol{v}\right)_{\mathcal{M}} + \epsilon^{-1}k\left(\boldsymbol{u}^\perp,\boldsymbol{v}\right)_{\mathcal{M}} - Fr^2 k\left(\boldsymbol{\eta},\nabla\cdot\boldsymbol{v}\right)_{\mathcal{A}} + k\left(\mathcal{B}\boldsymbol{u},\boldsymbol{v}\right) \\ &+ Fr^2\left(\boldsymbol{\eta},\boldsymbol{w}\right)_{\mathcal{A}} + Fr^2 k\left(\nabla\cdot\boldsymbol{u},\boldsymbol{w}\right)_{\mathcal{A}} \\ \le& \left(1+\tfrac{k}{\epsilon}\right)\|\boldsymbol{u}\|_{\mathcal{M}}\|\boldsymbol{v}\|_{\mathcal{M}} + Fr^2 k\|\boldsymbol{\eta}\|_{\mathcal{A}}\|\nabla\cdot\boldsymbol{v}\|_{\mathcal{A}} + k\|\mathcal{B}\boldsymbol{u}\|\|\boldsymbol{v}\| \\ &+ Fr^2\|\boldsymbol{\eta}\|_{\mathcal{A}}\|\boldsymbol{w}\|_{\mathcal{A}} + Fr^2 k\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}\|\|\boldsymbol{w}\|_{\mathcal{A}}.\end{aligned} \tag{3.9}$$

At this point, we use the boundedness of $\mathcal{B}$ assumed in (2.27) and the norm equivalence of $\|\cdot\|$ and $\|\cdot\|_{\mathcal{M}}$ in (2.26) to obtain

$$\begin{aligned}\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right) \le& \left(1+\tfrac{k}{\epsilon}+\tfrac{kB^*}{C_{\mathcal{M}}^2}\right)\|\boldsymbol{u}\|_{\mathcal{M}}\|\boldsymbol{v}\|_{\mathcal{M}} + Fr^2 k\|\boldsymbol{\eta}\|_{\mathcal{A}}\|\nabla\cdot\boldsymbol{v}\|_{\mathcal{A}} \\ &+ Fr^2\|\boldsymbol{\eta}\|_{\mathcal{A}}\|\boldsymbol{w}\|_{\mathcal{A}} + Fr^2 k\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}\|\|\boldsymbol{w}\|_{\mathcal{A}}.\end{aligned} \tag{3.10}$$

We can rewrite the right-hand side of this as the inner product of two vectors and apply the discrete Cauchy–Schwarz to bound this by

$$\begin{aligned}\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right) \le& \sqrt{\left(1+\tfrac{k}{\epsilon}+\frac{kB^*}{C_{\mathcal{M}}^2}\right)\|\boldsymbol{u}\|_{\mathcal{M}}^2 + 2Fr^2\|\boldsymbol{\eta}\|_{\mathcal{A}}^2 + Fr^2k^2\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2} \\ &\times\sqrt{\left(1+\tfrac{k}{\epsilon}+\frac{kB^*}{C_{\mathcal{M}}^2}\right)\|\boldsymbol{v}\|_{\mathcal{M}}^2 + 2Fr^2\|\boldsymbol{w}\|_{\mathcal{A}}^2 + Fr^2k^2\|\nabla\cdot\boldsymbol{v}\|_{\mathcal{A}}^2} \\ \le& C\left\|(\boldsymbol{u},\boldsymbol{\eta})\right\|_{\hat{b}}\left\|(\boldsymbol{v},\boldsymbol{w})\right\|_{\hat{b}}.\end{aligned} \tag{3.11}$$

□

If one includes the damping term in the inner product $\hat{b}$, the continuity estimate is independent of B^*. However, since the damping is typically small, it has little effect on preconditioner performance. Furthermore, omitting nonlinear damping from the preconditioner avoids the need to reassemble at each linear iteration.

THEOREM 3.2. *The bilinear form $\hat{a}$ is inf-sup stable with respect to the $\|\cdot\|_{\hat{b}}$ norm with constant no smaller than $\frac{1}{2\sqrt{3}}$.*

Proof. Let $(\boldsymbol{u},\boldsymbol{\eta}) \in \boldsymbol{V}_h^N \times \boldsymbol{W}_h^N$ be given and put $\boldsymbol{v}=\boldsymbol{u}$ and $\boldsymbol{w}=\boldsymbol{\eta}+k\nabla\cdot\boldsymbol{u}$. Then we see that

$$\begin{aligned}\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right) =& \left(\boldsymbol{u},\boldsymbol{u}\right)_{\mathcal{M}} + \tfrac{k}{\epsilon}\left(\boldsymbol{u}^\perp,\boldsymbol{u}\right)_{\mathcal{M}} - Fr^2 k\left(\boldsymbol{\eta},\nabla\cdot\boldsymbol{u}\right)_{\mathcal{A}} + k\left(\mathcal{B}\boldsymbol{u},\boldsymbol{u}\right) \\ &+ Fr^2\left(\boldsymbol{\eta},\boldsymbol{\eta}+k\nabla\cdot\boldsymbol{u}\right)_{\mathcal{A}} + Fr^2 k\left(\nabla\cdot\boldsymbol{u},\boldsymbol{\eta}+k\nabla\cdot\boldsymbol{u}\right)_{\mathcal{A}} \\ =& \|\boldsymbol{u}\|_{\mathcal{M}}^2 + k\left(\mathcal{B}\boldsymbol{u},\boldsymbol{u}\right) \\ &+ Fr^2\|\boldsymbol{\eta}\|_{\mathcal{A}}^2 + Fr^2 k\left(\boldsymbol{\eta},\nabla\cdot\boldsymbol{u}\right)_{\mathcal{A}} + Fr^2k^2\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2.\end{aligned} \tag{3.12}$$

Now, the semidefiniteness of $\mathcal{B}$ and standard estimates let us make the bound

$$\begin{aligned}\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right) \ge& \|\boldsymbol{u}\|_{\mathcal{M}}^2 + Fr^2\|\boldsymbol{\eta}\|_{\mathcal{A}}^2 + Fr^2k^2\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2 \\ &- \tfrac{Fr^2}{2}\|\boldsymbol{\eta}\|_{\mathcal{A}}^2 - \tfrac{Fr^2k^2}{2}\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2 \\ =& \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \tfrac{Fr^2}{2}\|\boldsymbol{\eta}\|_{\mathcal{A}}^2 + \tfrac{Fr^2k^2}{2}\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2 \\ \ge& \tfrac{1}{2}\left\|(\boldsymbol{u},\boldsymbol{\eta})\right\|_{\hat{b}}^2.\end{aligned} \tag{3.13}$$

Now, we also have

$$\begin{aligned}\left\|(\boldsymbol{v},\boldsymbol{w})\right\|_{\hat{b}}^2 =& \|\boldsymbol{u}\|_{\mathcal{M}}^2 + Fr^2k^2\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2 + Fr^2\|\boldsymbol{\eta}+k\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2 \\ \le& \|\boldsymbol{u}\|_{\mathcal{M}}^2 + Fr^2k^2\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2 + 2Fr^2\left(\|\boldsymbol{\eta}\|_{\mathcal{A}}^2 + k^2\|\nabla\cdot\boldsymbol{u}\|_{\mathcal{A}}^2\right) \\ \le& 3\left\|(\boldsymbol{u},\boldsymbol{\eta})\right\|_{\hat{b}}^2.\end{aligned} \tag{3.14}$$

Hence,

$$\hat{a}\left((\boldsymbol{u},\boldsymbol{\eta}),(\boldsymbol{v},\boldsymbol{w})\right) \geq \tfrac{1}{2}\left\|(\boldsymbol{u},\boldsymbol{\eta})\right\|\left\|(\boldsymbol{v},\boldsymbol{w})\right\| \geq \frac{1}{2\sqrt{3}}\left\|(\boldsymbol{u},\boldsymbol{\eta})\right\|_{\hat{b}}\left\|(\boldsymbol{v},\boldsymbol{w})\right\|_{\hat{b}}, \tag{3.15}$$

and the result follows. □

4. More about $\mathcal{A}$.

The major cost of applying our block diagonal preconditioner is the inversion of the upper-left block of (3.1):

$$\mathrm{C} = \mathrm{M}^{V} + Fr^2k^2\mathrm{E}^{\mathcal{A}}. \tag{4.1}$$

PROPOSITION 4.1. *If the densities are positive and strictly increasing (that is, if $0 < \rho_1 < \rho_2 < \ldots < \rho_N$), then A is positive-definite.*

Proof. Clearly, the result holds if $N = 1$. For $N > 1$, consider taking one step of Gaussian elimination on A, by which

$$\begin{aligned}\begin{bmatrix}\rho_1 & \rho_1 & \rho_1 & \cdots & \rho_1\\ \rho_1 & \rho_2 & \rho_2 & \cdots & \rho_2\\ & \vdots & & \ddots & \\ \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_N\end{bmatrix} &\to \begin{bmatrix}\rho_1 & \rho_1 & \rho_1 & \cdots & \rho_1\\ 0 & \rho_2-\rho_1 & \rho_2-\rho_1 & \cdots & \rho_2-\rho_1\\ & \vdots & & \ddots & \\ 0 & \rho_2-\rho_1 & \rho_3-\rho_1 & \cdots & \rho_N-\rho_1\end{bmatrix}\\ &\equiv \begin{bmatrix}\rho_1 & \rho_1 & \rho_1 & \cdots & \rho_1\\ 0 & \rho_{1,2} & \rho_{1,2} & \cdots & \rho_{1,2}\\ & \vdots & & \ddots & \\ 0 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,N}\end{bmatrix}.\end{aligned} \tag{4.2}$$

Under our assumptions on ρ_i, the sequence $\{\rho_{1,i+1}\}_{i=1}^{N-1}$ is positive and strictly increasing. Hence, Gaussian elimination without interchanges continues with positive pivots, which is equivalent to positive-definiteness [36]. □

Now, we give an explicit formula for the inverse of $\mathcal{A}$ and estimates on its extremal eigenvalues. This sets us up to discuss preconditioners for C in the following section.

4.1. An explicit inverse for $\mathcal{A}$.

PROPOSITION 4.2. *Define the matrix $\mathcal{C}$ to be the $N \times N$ symmetric tridiagonal matrix with*

$$\mathcal{C}_{ii} = \begin{cases}\frac{1}{\rho_1} + \frac{1}{\rho_2-\rho_1}, & i = 1,\\ \frac{1}{\rho_i-\rho_{i-1}} + \frac{1}{\rho_{i+1}-\rho_i}, & 2 < i < N-1,\\ \frac{1}{\rho_n-\rho_{n-1}}, & i = N,\end{cases} \tag{4.3}$$

and off-diagonal entries

$$\mathcal{C}_{i,i+1} = \mathcal{C}_{i+1,i} = -\frac{1}{\rho_{i+1}-\rho_i}, \quad 1 \leq i \leq N-1. \tag{4.4}$$

Then $\mathcal{C}$ is the inverse of $\mathcal{A}$ given in (2.20).

Proof. The result can be obtained by Gauss–Jordan elimination on $\mathcal{A}$, although the notation for the case of general N is quite cumbersome. Here, we confirm the result by verifying $\mathcal{CA} = I$.

Since the diagonal of $\mathcal{C}$ is defined piecewise, we proceed in a few cases. Consider the first row of $\mathcal{S} = \mathcal{CA}$:

$$\begin{aligned}\mathcal{S}_{11} &= \mathcal{C}_{11}\mathcal{A}_{11} + \mathcal{C}_{12}\mathcal{A}_{21}\\ &= \left(\frac{1}{\rho_1} + \frac{1}{\rho_2-\rho_1}\right)\rho_1 - \frac{1}{\rho_2-\rho_1}\rho_1 = 1.\end{aligned} \tag{4.5}$$

For any $j > 1$, we have that $\mathcal{A}_{1j} = \rho_1$ and $\mathcal{A}_{2j} = \rho_2$, so

$$\begin{aligned} \mathcal{S}_{1j} &= \mathcal{C}_{11}\mathcal{A}_{1j} + \mathcal{C}_{12}\mathcal{A}_{2j} \\ &= \left(\frac{1}{\rho_1} + \frac{1}{\rho_2 - \rho_1}\right)\rho_1 - \frac{1}{\rho_2 - \rho_1}\rho_2 \\ &= 1 - \frac{\rho_2 - \rho_1}{\rho_2 - \rho_1} = 0. \end{aligned} \tag{4.6}$$

Now, for $2 \le i < N$, we have

$$\begin{aligned} \mathcal{S}_{ii} &= \mathcal{C}_{i,i-1}\mathcal{A}_{i-1,i} + \mathcal{C}_{i,i}\mathcal{A}_{i,i} + \mathcal{C}_{i,i+1}\mathcal{A}_{i+1,i} \\ &= -\frac{1}{\rho_i - \rho_{i-1}}\rho_{i-1} + \left(\frac{1}{\rho_i - \rho_{i-1}} + \frac{1}{\rho_{i+1} - \rho_i}\right)\rho_i - \frac{1}{\rho_{i+1} - \rho_i}\rho_i = 1. \end{aligned} \tag{4.7}$$

For $j > i$, we have

$$\begin{aligned} \mathcal{S}_{ij} &= \mathcal{C}_{i,i-1}\mathcal{A}_{i-1,j} + \mathcal{C}_{i,i}\mathcal{A}_{i,j} + \mathcal{C}_{i,i+1}\mathcal{A}_{i+1,j} \\ &= -\frac{1}{\rho_i - \rho_{i-1}}\rho_{i-1} + \left(\frac{1}{\rho_i - \rho_{i-1}} + \frac{1}{\rho_{i+1} - \rho_i}\right)\rho_i - \frac{1}{\rho_{i+1} - \rho_i}\rho_{i+1} = 0, \end{aligned} \tag{4.8}$$

and for $j < i$,

$$\begin{aligned} \mathcal{S}_{ij} &= \mathcal{C}_{i,i-1}\mathcal{A}_{i-1,j} + \mathcal{C}_{i,i}\mathcal{A}_{i,j} + \mathcal{C}_{i,i+1}\mathcal{A}_{i+1,j} \\ &= -\frac{1}{\rho_i - \rho_{i-1}}\rho_j + \left(\frac{1}{\rho_i - \rho_{i-1}} + \frac{1}{\rho_{i+1} - \rho_i}\right)\rho_j - \frac{1}{\rho_{i+1} - \rho_i}\rho_j = 0. \end{aligned} \tag{4.9}$$

Finally, we handle the last row. The diagonal entry there is

$$\begin{aligned} \mathcal{S}_{N,N} &= \mathcal{C}_{N,N-1}\mathcal{A}_{N-1,N} + \mathcal{C}_{N,N}\mathcal{A}_{N,N} \\ &= -\frac{1}{\rho_N - \rho_{N-1}}\rho_{N-1} + \frac{1}{\rho_N - \rho_{N-1}}\rho_N = 1, \end{aligned} \tag{4.10}$$

and for any $1 \le j < N$,

$$\begin{aligned} \mathcal{S}_{N,N} &= \mathcal{C}_{N,N-1}\mathcal{A}_{N-1,j} + \mathcal{C}_{N,N}\mathcal{A}_{N,j} \\ &= -\frac{1}{\rho_N - \rho_{N-1}}\rho_j + \frac{1}{\rho_N - \rho_{N-1}}\rho_j = 0. \end{aligned} \tag{4.11}$$

□

Finally, since $\mathcal{C}$ is tridiagonal and symmetric positive-definite, it has a factorization

$$\mathcal{C} = \mathcal{L}\mathcal{D}\mathcal{L}^T \tag{4.12}$$

with bidiagonal $\mathcal{L}$ and diagonal $\mathcal{D}$ with positive entries.

4.2. The spectrum of $\mathcal{A}$. Subsequent analysis will rely on knowing things about the spectrum of $\mathcal{A}$, and we are able to give certain instructive spectral bounds here. Since $\mathcal{A}$ is symmetric and positive-definite, we let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N > 0$ be its eigenvalues, arranged in nonincreasing order.

PROPOSITION 4.3. *The largest eigenvalue of $\mathcal{A}$ satisfies*

$$N\rho_1 \le \lambda_1 \le \sum_{j=1}^{N} \rho_j. \tag{4.13}$$

Proof. We handle the upper bound by Gerschgorin's circle theorem. Owing to the structure of $\mathcal{A}$, the largest outer extent of a Gerschgorin disk comes from the final row, and the maximal value is

$$\rho_N + \sum_{j=1}^{N-1} \rho_j = \sum_{i=1}^{N} \rho_i. \tag{4.14}$$

Now, we derive the lower bound in (4.13), which confirms that λ_1 is in fact comparable to N. Since λ_1 maximizes the Rayleigh quotient,

$$\lambda_1 = \max_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}}, \tag{4.15}$$

using any particular choice of nonzero $\mathbf{x}$ in the Rayleigh quotient gives a lower bound for λ_1. We chose the vector $\mathbf{x}$ consisting entire of ones. Since $\mathbf{x}^T \mathbf{x} = N$,

$$N\lambda_1 \geq \mathbf{x}^T \mathcal{A} \mathbf{x}. \tag{4.16}$$

Proceeding, the entries of $\mathcal{A}\mathbf{x}$ are just the row sums of $\mathcal{A}$:

$$(\mathcal{A}\mathbf{x})_i = \sum_{j=1}^{i-1} \rho_j + \sum_{j=i}^{N} \rho_i = \left(\sum_{j=1}^{i-1} \rho_j \right) + (N - i + 1)\, \rho_i. \tag{4.17}$$

Evaluating $\mathbf{x}^T \mathcal{A} \mathbf{x}$ gives

$$\begin{aligned} \mathbf{x}^T \mathcal{A} \mathbf{x} &= \sum_{i=1}^{N} (\mathcal{A}\mathbf{x})_i = \sum_{i=1}^{N} \left[\left(\sum_{j=1}^{i-1} \rho_j \right) + (N - i + 1)\, \rho_i \right] \\ &= \sum_{i=1}^{N} (N - i)\, \rho_i + \sum_{i=1}^{N} (N - i + 1)\, \rho_i = \sum_{i=1}^{N} (2N - 2i + 1)\, \rho_i. \end{aligned} \tag{4.18}$$

Since $\rho_1 < \rho_i$ for $i > 1$,

$$N\lambda_1 \geq \rho_1 \sum_{i=1}^{N} (2N - 2i + 1) = \rho_1 \left[2N^2 - 2\frac{N(N+1)}{2} + N \right] = N^2 \rho_1. \tag{4.19}$$

This proves the lower bound. □

Similar techniques can lead to upper and lower bounds on the minimal eigenvalue λ_N.

THEOREM 4.4. *Let* $\delta\rho_* = \min_{1 \leq i \leq N-1} \rho_{i+1} - \rho_i$ *and* $\delta\rho^* = \max_{1 \leq i \leq N_1} \rho_{i+1} - \rho_i$. *Then*

$$\tfrac{\delta\rho_*}{4} \leq \lambda_N \leq \tfrac{3\delta\rho^*}{10}. \tag{4.20}$$

Proof. The Gerschgorin t Theorem again bounds the maximal eigenvalue of $\mathcal{C}$, which is the reciprocal of the minimal eigenvalue of $\mathcal{A}$. Consider the first row of $\mathcal{C}$. The diagonal plus sum of magnitudes of off-diagonal entries yields

$$\tfrac{1}{\rho_1} + \tfrac{2}{\rho_2 - \rho_1} = \tfrac{\rho_1 + \rho_2}{\rho_1(\rho_2 - \rho_1)}. \tag{4.21}$$

We then use (2.1) to bound this by

$$\tfrac{3}{\rho_2-\rho_1} \le \tfrac{3}{\delta\rho_*}.$$

Then, for $2 \le i \le N-1$, the diagonal plus sum of off-diagonal magnitudes gives

$$2\left[\tfrac{1}{\rho_i-\rho_i-1} + \tfrac{1}{\rho_{i+1}-\rho_i}\right] \le \tfrac{4}{\delta\rho_*}. \tag{4.22}$$

Finally, the outer limit of the Gerschgorin disk for the final row is

$$\tfrac{2}{\rho_N-\rho_{N-1}} \le \tfrac{2}{\delta\rho_*}. \tag{4.23}$$

Taking the maximum over these three calculations gives that

$$\tfrac{1}{\lambda_N} \le \tfrac{4}{\delta\rho_*}, \tag{4.24}$$

and the reciprocal of this inequality gives the lower bound.

To establish the upper bound, we again consider the Rayleigh quotient on a particular vector. Pick some vector $\mathbf{x}$ such that for a fixed $3 \le i \le N-2$

$$\mathbf{x}_j = \begin{cases} 1, & j = i, \\ -1, & |j-i| = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.25}$$

Since $\mathbf{x}$ is nonzero only in entries $i-1, i, i+1$, we directly compute the relevant entries of $\mathcal{C}\mathbf{x}$ using (4.3) and (4.4):

$$\begin{aligned} (\mathcal{C}\mathbf{x})_{i-1} &= \sum_{j=1}^{N} \mathcal{C}_{i-1,j}\mathbf{x}_j = \mathcal{C}_{i-1,i-1}\mathbf{x}_{i-1} + \mathcal{C}_{i-1,i}\mathbf{x}_i \\ &= -\left(\tfrac{1}{\rho_{i-1}-\rho_{i-2}} + \tfrac{1}{\rho_i-\rho_{i-1}}\right) - \tfrac{1}{\rho_i-\rho_{i-1}} \\ &= -\tfrac{1}{\rho_{i-1}-\rho_{i-2}} - \tfrac{2}{\rho_i-\rho_{i-1}}. \end{aligned} \tag{4.26}$$

$$\begin{aligned} (\mathcal{C}\mathbf{x})_i &= \sum_{j=1}^{N} \mathcal{C}_{i,j}\mathbf{x}_j = \mathcal{C}_{i,i-1}\mathbf{x}_{i-1} + \mathcal{C}_{i,i}\mathbf{x}_i + \mathcal{C}_{i,i+1}\mathbf{x}_{i+1} \\ &= \tfrac{1}{\rho_i-\rho_{i-1}} + \left(\tfrac{1}{\rho_i-\rho_{i-1}} + \tfrac{1}{\rho_{i+1}-\rho_i}\right) + \tfrac{1}{\rho_{i+1}-\rho_i} \\ &= \tfrac{2}{\rho_i-\rho_{i-1}} + \tfrac{2}{\rho_{i+1}-\rho_i}. \end{aligned} \tag{4.27}$$

Similarly, we can compute

$$(\mathcal{C}\mathbf{x})_{i+1} = -\tfrac{2}{\rho_{i+1}-\rho_i} - \tfrac{1}{\rho_{i+2}-\rho_{i+1}}. \tag{4.28}$$

Now, we use the results to directly calculate that

$$\mathbf{x}^T\mathcal{C}\mathbf{x} = \tfrac{1}{\rho_{i-1}-\rho_{i-2}} + \tfrac{4}{\rho_i-\rho_{i-1}} + \tfrac{4}{\rho_{i+1}-\rho_i} + \tfrac{1}{\rho_{i+2}-\rho_{i+1}} \ge \tfrac{10}{\delta\rho^*}. \tag{4.29}$$

Now, we note that $\mathbf{x}^T\mathbf{x} = 3$ for this $\mathbf{x}$ and using the Rayleigh quotient gives the upper bound on λ_N. ∎

Assuming some kind of comparability between $\delta\rho_*$ and $\delta\rho^*$, both are on the order of N. This gives a spectral condition number (ratio of extremal eigenvalues) for $\mathcal{A}$ on the order of N^2.

5. Simplifying the preconditioner.

Our weighted-norm preconditioner (3.1) provides parameter robustness, but also maintains an all-to-all coupling between the layers that can become expensive as the number of layers increases. In this section, we propose two approaches to overcoming this difficulty. In the first case, we simply ignore the interlayer coupling. We are able to prove that this strategy is more effective than the the N^2 conditioning of A might otherwise suggest. In the second case, we make use of the special properties of $\mathcal{A}$ derived above to propose a change of variables in the upper-left block of (3.1) that renders coupling only between adjacent layers.

5.1. Neglecting interlayer coupling.

The bilinear form

$$c(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u},\boldsymbol{v})_{\mathcal{M}} + Fr^2k^2\left(\nabla\cdot\boldsymbol{u},\nabla\cdot\boldsymbol{v}\right)_{\mathcal{A}} \tag{5.1}$$

yields the matrix (4.1) under discretization, and we want to compare c to the simpler form obtained by replacing the $\mathcal{A}$-weighted inner product with the standard one:

$$\hat{c}(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u},\boldsymbol{v})_{\mathcal{M}} + Fr^2k^2\left(\nabla\cdot\boldsymbol{u},\nabla\cdot\boldsymbol{v}\right). \tag{5.2}$$

The latter form gives rise to the block diagonal matrix

$$\hat{\mathrm{C}} = \mathrm{M}^V + Fr^2k^2\mathrm{E}, \tag{5.3}$$

which we can consider using as a preconditioner for the matrix derived from $c(\cdot,\cdot)$. Both c and $\hat{c}$ are symmetric and positive-definite, and showing an equivalence controls eigenvalues of the system obtained by preconditioning one with the other.

As a first attempt, $\mathcal{A}$ is symmetric and positive-definite, and we can use the Rayleigh quotient pointwise inside of integrals to obtain:

$$\|\boldsymbol{w}\|_{\mathcal{A}}^2 = \int_\Omega (\mathcal{A}\boldsymbol{w})\cdot\boldsymbol{w}\,dx \geq \int_\Omega \lambda_1\left(\boldsymbol{w}\cdot\boldsymbol{w}\right) = \lambda_1\|\boldsymbol{w}\|^2 \tag{5.4}$$

with a similar upper bound of $\|\boldsymbol{w}\|_{\mathcal{A}}^2 \leq \lambda_1\|\boldsymbol{w}\|^2$.

Using this observation,

$$\lambda_N\left(\|\boldsymbol{u}\|_{\mathcal{M}}^2 + k^2Fr^2\|\nabla\cdot\boldsymbol{u}\|^2\right) \leq c(\boldsymbol{u},\boldsymbol{u}) \leq \lambda_1\left(\|\boldsymbol{u}\|_{\mathcal{M}}^2 + k^2Fr^2\|\nabla\cdot\boldsymbol{u}\|^2\right), \tag{5.5}$$

so that an equivalence between c and $\hat{c}$ holds with a condition number of λ_1/λ_N, which is quadratic in the number of layers. With more careful consideration, however, we are able to prove a tighter bound.

In this analysis, we will make the *inverse assumption* that there exists some $C_I > 0$, independent of $\boldsymbol{u}$ and h such that

$$\|\nabla\cdot\boldsymbol{u}\| \leq \tfrac{C_I}{h}\|\boldsymbol{u}\|_{\mathcal{M}} \tag{5.6}$$

holds for all $\boldsymbol{u}\in\boldsymbol{V}_h^N$ with some $C_I > 0$ independent of $\boldsymbol{u}$. This estimate is a theorem for standard H^1 polynomial spaces [7] and is a commonly made assumption for $H(\mathrm{div})$ spaces. In our case, it follows from the standard $H(\mathrm{div})$ inverse assumption in each component plus the equivalence of $\|\cdot\|_{\mathcal{M}}$ to the N-way L^2 inner product.

To simplify our notation, we introduce the quantity

$$q = C_IkFr. \tag{5.7}$$

THEOREM 5.1. *For all* $\mathbf{u} \in \boldsymbol{V}_h^N$, *the equivalence*

$$\chi_0 \hat{c}(\mathbf{u}, \mathbf{u}) \le c(\mathbf{u}, \mathbf{u}) \le \chi_1 \hat{c}(\mathbf{u}, \mathbf{u}) \tag{5.8}$$

holds, where

$$\chi_0 = \frac{\lambda_N q^2 + h^2}{q^2 + h^2}, \quad \chi_1 = \frac{\lambda_1 q^2 + h^2}{q^2 + h^2}. \tag{5.9}$$

Proof. We first prove the upper bound involving χ_1, applying the Rayleigh quotient for $\mathcal{A}$ pointwise to obtain

$$c(\boldsymbol{u}, \boldsymbol{u}) \le \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \lambda_1 Fr^2 k^2 \|\nabla \cdot \boldsymbol{u}\|^2. \tag{5.10}$$

Next, for some $0 \le \alpha \le 1$ to be specified, we split the $\|\nabla \cdot \boldsymbol{u}\|^2$ term

$$c(\boldsymbol{u}, \boldsymbol{v}) \le \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \alpha \lambda_1 Fr^2 k^2 \|\nabla \cdot \boldsymbol{u}\|^2 + (1-\alpha) \lambda_1 Fr^2 k^2 \|\nabla \cdot \boldsymbol{u}\|^2, \tag{5.11}$$

and using the inverse assumption (5.6), we have

$$c(\boldsymbol{u}, \boldsymbol{v}) \le \left(1 + \tfrac{\alpha \lambda_1 q^2}{h^2}\right) \|\boldsymbol{u}\|_{\mathcal{M}}^2 + (1-\alpha)\, \lambda_1 Fr^2 k^2 \|\nabla \cdot \boldsymbol{u}\|^2. \tag{5.12}$$

The best bound here will be obtained if we choose α to equalize the coefficients of the terms appearing in the bilinear form, or

$$1 + \tfrac{\alpha \lambda_1 q^2}{h^2} = (1-\alpha)\, \lambda_1.$$

This is readily solved to find

$$\alpha = \frac{(\lambda_1 - 1)\, h^2}{\lambda_1 \left(h^2 + q^2\right)}. \tag{5.13}$$

The coefficient of $k^2 Fr^2 \|\nabla \cdot \boldsymbol{u}\|^2$ in our estimate is $\alpha_1 \lambda_1$, which is equal to the claimed value χ_1. The coefficient of $\|\boldsymbol{u}\|_{\mathcal{M}}^2$ has the same value, completing the upper bound.

Now, we consider the lower bound, which begins in the same way, using the lower bound on the Rayleigh quotient to write

$$c(\boldsymbol{u}, \boldsymbol{v}) \ge \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \lambda_N Fr^2 k^2 \|\nabla \cdot \boldsymbol{u}\|^2. \tag{5.14}$$

Now, we additively split the L^2 term with some $0 < \alpha < 1$:

$$c(\boldsymbol{u}, \boldsymbol{u}) \ge (1-\alpha) \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \alpha \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \lambda_N k^2 Fr^2 \|\nabla \cdot \boldsymbol{u}\|^2. \tag{5.15}$$

Now, we rearrange the inverse assumption to bound $\|\boldsymbol{u}\|_{\mathcal{M}}$ below by $\frac{h}{C_I} \|\nabla \cdot \boldsymbol{u}\|$:

$$\begin{aligned} c(\boldsymbol{u}, \boldsymbol{u}) &\ge (1-\alpha) \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \left(\tfrac{\alpha h^2}{C_I^2} + k^2 Fr^2 \lambda_N\right) \|\nabla \cdot \boldsymbol{u}\|^2 \\ &= (1-\alpha) \|\boldsymbol{u}\|_{\mathcal{M}}^2 + \left(\tfrac{\alpha h^2}{q^2} + \lambda_N\right) k^2 Fr^2 \|\nabla \cdot \boldsymbol{u}\|^2 . \end{aligned} \tag{5.16}$$

Again, the optimal choice of α will balance the coefficients, so we solve

$$1 - \alpha = \tfrac{\alpha h^2}{q^2} + \lambda_N \quad \to \quad \alpha = \frac{q^2 \left(1 - \lambda_N\right)}{q^2 + h^2}, \tag{5.17}$$

so that $1 - \alpha = \chi_0$ as claimed. □

This theorem shows a somewhat complex relationship between the physical and discretization parameters and the equivalence bounds obtained by neglecting the interlayer coupling. The lower bound is somewhat simpler to unpack. Since $\lambda_N > 0$ but decays like $1/N$, we always have

$$\chi_0 \geq \frac{h^2}{q^2 + h^2},$$

which is *independent* of the number of layers. Fixing h and letting q (here, a proxy for the time step) become small presents no problems. On the other hand, keeping a nondegenerate lower bound when $h \to 0$ also requires $q \to 0$ at a comparable rate.

The asymptotics of the upper bound are a bit different. We have that $\lambda_1 = \mathcal{O}(N)$ as we increase the number of layers. However, this only makes $\chi_1/\chi_0 = \mathcal{O}(N)$ rather than the naive $\mathcal{O}(N)^2$ posited initially. Also, for a fixed number of layers, two comments are in order. First, we always have $\chi_1 < \lambda_1$. Second, we can decrease the effect of large λ_1 by reducing the time step relative to the mesh size, for

$$\chi_1 = \frac{\lambda_1 q^2 + h^2}{q^2 + h^2} = \frac{\lambda_1 \left(\frac{q}{h}\right)^2 + 1}{\left(\frac{q}{h}\right)^2 + 1}.$$

5.2. A block tridiagonal reformulation. Neglecting the interlayer coupling in our preconditioner is better than initially thought, and performs well for practical numbers of layers. Here, we sketch an alternate approach that should also sparsify the preconditioner while maintaining the layer independence. This approach relies heavily on the tridiagonal inverse of the coupling matrix $\mathcal{A}$.

For the bilinear form c from (5.1) and some bounded linear functional $f \in (\boldsymbol{V}_h^N)'$, consider the variational problem

$$c(\boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{v}), \quad \boldsymbol{v} \in V_h^N. \tag{5.18}$$

Using (4.12) in this, we write

$$\mathcal{A} = \mathcal{C}^{-1} = \left(\mathcal{L}\mathcal{D}\mathcal{L}^T\right)^{-1} = \mathcal{L}^{-T}\mathcal{D}^{-1}\mathcal{L}^{-1}, \tag{5.19}$$

so that

$$\begin{aligned} c(\boldsymbol{u}, \boldsymbol{v}) &= (\mathcal{M}\boldsymbol{u}, \boldsymbol{v}) + Fr^2k^2(\mathcal{L}^{-T}\mathcal{D}^{-1}\mathcal{L}^{-1}\nabla\cdot\boldsymbol{u}, \nabla\cdot\boldsymbol{v}) \\ &= (\mathcal{M}\boldsymbol{u}, \boldsymbol{v}) + Fr^2k^2(\mathcal{D}^{-1}\nabla\cdot\mathcal{L}^{-1}\boldsymbol{u}, \nabla\cdot\mathcal{L}^{-1}\boldsymbol{v}). \end{aligned} \tag{5.20}$$

We introduce auxiliary variables $\widetilde{\boldsymbol{u}} = \mathcal{L}^{-1}\boldsymbol{u}$ and $\widetilde{\boldsymbol{v}} = \mathcal{L}^{-1}\boldsymbol{v}$ and define $\widetilde{\mathcal{C}} = \mathcal{L}^T\mathcal{M}\mathcal{L}$. This varies spatially, but is pointwise tridiagonal. With these substitutions, we have

$$\begin{aligned} c(\boldsymbol{u}, \boldsymbol{v}) &= (\mathcal{M}\mathcal{L}\widetilde{\boldsymbol{u}}, \mathcal{L}\widetilde{\boldsymbol{v}}) + Fr^2k^2(\mathcal{D}^{-1}\nabla\cdot\widetilde{\boldsymbol{u}}, \nabla\cdot\widetilde{\boldsymbol{v}}) \\ &= (\widetilde{\mathcal{C}}\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) + Fr^2k^2(\mathcal{D}^{-1}\nabla\cdot\widetilde{\boldsymbol{u}}, \nabla\cdot\widetilde{\boldsymbol{v}}) \\ &\equiv c(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}). \end{aligned} \tag{5.21}$$

Now, we can write (5.18) as

$$\tilde{c}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) = \tilde{f}(\widetilde{\boldsymbol{v}}). \tag{5.22}$$

Hence, one could change variables and solve a sparser system, in which only adjacent layers are coupled through the tridiagonal matrix $\widetilde{\mathcal{C}}$, although this requires considerable care in the implementation.

6. Numerical results.

We have implemented a mixed finite element discretization of the tide model and developed all of our preconditioners within the Firedrake framework [30]. Firedrake is an automated system for the solution of PDEs using the finite element method. It generates efficient low-level code from the Unified Form Language in Python [1], and interfaces tightly with PETSc for scalable algebraic solvers. Firedrake also has a rich ability to interoperate with and extend PETSc [22], which facilitates the definition of auxiliary bilinear forms needed for weighted-norm preconditioning. Moreover, a facility to generate Runge–Kutta methods from a semidiscrete formulation was recently added to Firedrake through the Irksome project [16], and we use this to obtain the implicit midpoint rule.

Our numerical experiments primarily consist of testing preconditioners as a function of discretization and physical parameters. We discretize the problem on the unit square by taking an $N_x \times N_y$ mesh subdivided into right triangles and use lowest-order Raviart–Thomas elements for $\boldsymbol{u}_h$ and piecewise constants for $\boldsymbol{\eta}_h$. In all our cases, we solve the resulting linear systems using unrestarted GMRES with right preconditioning. We chose the right-hand side by choosing an initial condition for the IBVP at rest but for a small disturbance in the top layer and taking one step of the implicit midpoint rule using Irksome. We iterated to the PETSc default relative tolerance of 10^{-5}, which is appropriate for the low-order time and space discretizations under consideration. In certain cases, we found it necessary to use modified Gram–Schmidt orthogonalization, and so we used it throughout. Our techniques are not particular to the Raviart–Thomas elements or triangles. Much as in [21], we have also performed our experiments on rectangular Raviart–Thomas elements and trimmed serendipity elements [14, 18] with very similar results.

As a point of reference, we will compare the weighted-norm preconditioners under consideration to an incomplete LU factorization method [33] with no fill. (Firedrake natively stores the momentum and elevation variables separately, but PETSc performs nested dissection to reorder the unknowns before performing the factorizations.) For wave like equations with a reasonable time step and moderate physical parameters, this is not a terrible approach. We refer to the two plots in Figure 6.1. Both plots fix 5 layers with equidistributed densities between 1.03 and 1.06 and Rossby number $\epsilon = 1$. In the first plot, we vary the Froude number and in the second, we vary the CFL number $\Delta t/N$. In both cases, we have mesh independence, but we see a wide range of variation with respect to the physical and discretization parameters.

We repeat these same experiments, now with the weighted-norm preconditioner we proposed in (3.1). Applying this preconditioner requires at least approximately inverting the block diagonal matrix. The best (in terms of iteration count) we can hope for is obtained if those blocks are in fact inverted exactly. The bottom right block is diagonal for lowest-order elements and hence trivial to invert. For the top left block, we compute a sparse LU factorization in a setup phase and perform solves with this at each iteration. We can compare Figure 6.2 to those in 6.1 and see the potential benefit of our new preconditioner. Although we see some variation with respect to the Froude and CFL numbers, we seem to approach a relatively small and mesh-independent bound, even for rather extreme parameter values.

However, for scaling to very large problems, it is important to consider ways of bypassing sparse factorizations. A simple strategy for this is to replace the inversion of the top left block with a simple ILU(0) factorization, and we repeat the experiments from Figures 6.1 and 6.2 using this choice in Figure 6.3. As expected, we lose some parameter robustness, but this could still give a practical result. These plots show

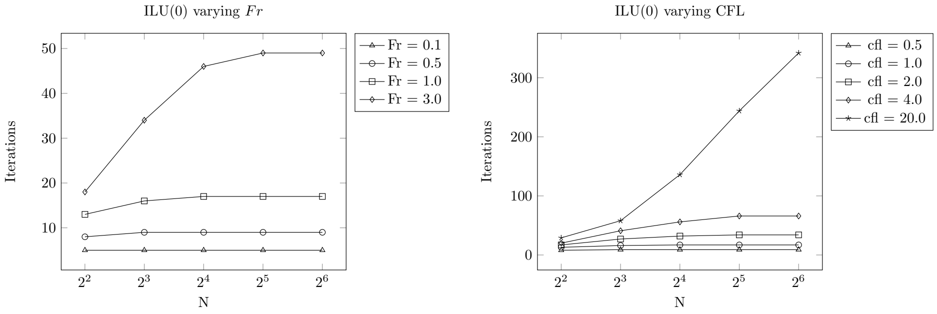

FIG. 6.1. *Performance of $ILU(0)$ preconditioner as a function of mesh parameters for various Froude and CFL numbers. Throughout, we fix $\epsilon = 1$ and consider 5 layers with densities varying between 1.03 and 1.06. We see eventual mesh independence, but the number of iterations varies considerably with fixing the CFL number as 1 and varying the Froude number (left) or fixing $Fr = 1$ and varying the CFL number (right).*

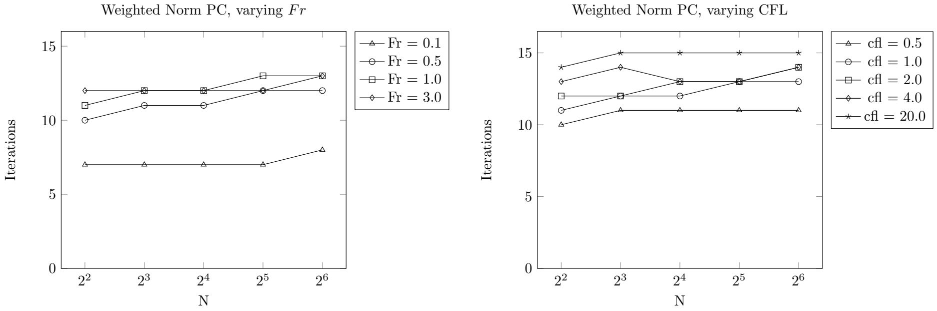

FIG. 6.2. *Performance of the preconditioner* (3.1) *using exact inversion of the blocks. Parameters are the same as in Figure* 6.1*.*

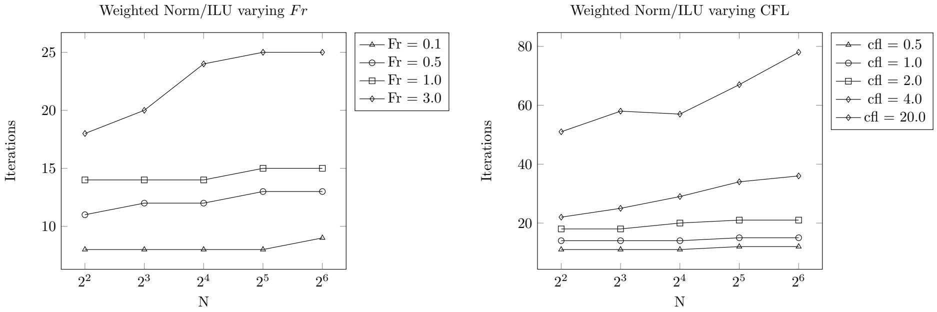

FIG. 6.3. *Repeating the experiments in Figure* 6.2*, but with the inverse of the top left block approximated by $ILU(0)$. We see an increase in iteration count, and greater parameter dependence. At moderate parameter values the increased iteration count is relatively small.*

that even running at CFL number requires only about 20 iterations per time step, and ILU(0) costs about as much as a matrix-vector product to apply. We remark that some adaptation of an $H(\mathrm{div})$ multigrid [3] inside the weighted-norm preconditioner plausibly would recover parameter robustness. However, these methods require smoothers that couple all degrees of freedom around vertex patches, hence will scale poorly with respect to the number of layers.

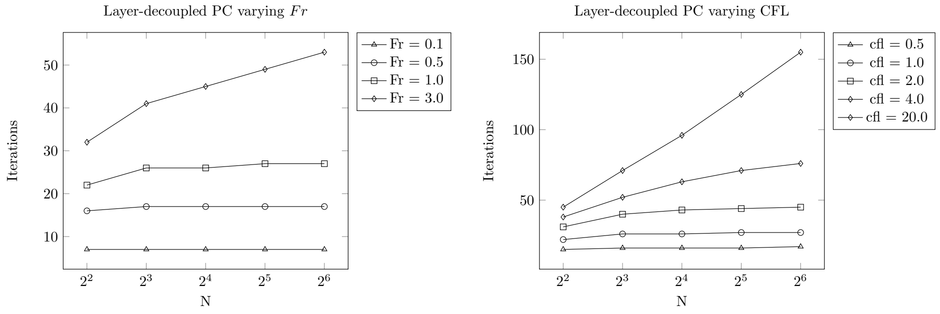

FIG. 6.4. *Performance of the weighted-norm preconditioner using the simplified form* (5.2) *in the top left block. Exact inversion of the blocks.*

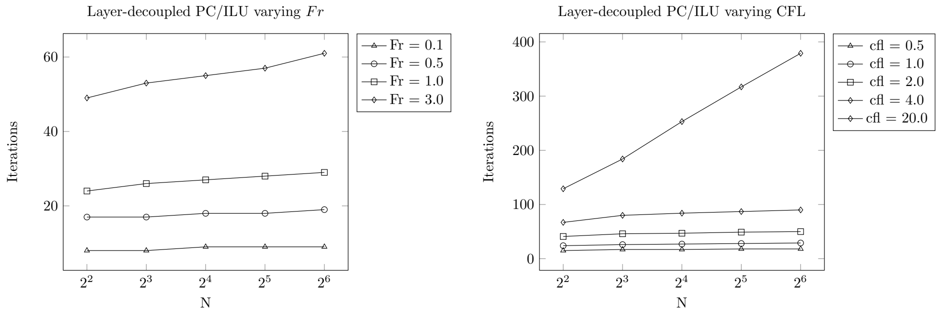

FIG. 6.5. *Performance of decoupled preconditioner with $ILU(0)$ for the top left block.*

We also repeat these experiments using the decoupled preconditioner suggested in (5.2) in the upper left block. Again, we present iteration counts exactly inverting the blocks in Figure 6.4 and using ILU(0) in the top left block in Figure 6.5. Perhaps unsurprisingly, we lose some parameter robustness with respect to the Froude and CFL numbers, but our iteration counts are only about 2–4 times as large as the respective iteration counts in Figures 6.2 and 6.3. The much-reduced sparsity of the preconditioner and hence its ILU(0) factorization could compensate for that increase.

In addition to the scaling performance we have considered, we also computed the eigenvalues of the preconditioned system. Here, we use the lowest-order Raviart–Thomas elements on an $N \times N$ triangular mesh, fix the time step as $1/N$, take $Fr = 1 = \epsilon$, and consider 4 fluid layers. The eigenvalues of the original matrix and the preconditioned system, with and without layer coupling, are presented in Figure 6.6. Although the eigenvalues of the original system are real, the small eigenvalues and lack of clustering create ill-conditioning. With the layer-coupled preconditioner (3.1), we introduce complex eigenvalues, but these are bounded away from the origin and are better clustered. The preconditioner (5.1) seems to give a kind of hybrid eigenvalue clustering, with some along the real axis as well as a larger ring of complex eigenvalues.

Now, we want to comment on the dependence of the preconditioners as a function of the number of layers. For this, we fixed an $N \times N$ mesh with $N = 64$ divided into triangles, fixed $Fr = \epsilon = 1$ and $dt = 2/N = 0.03125$, and considered the number of iterations required to solve the linear system with various preconditioners—ILU on the original system and preconditioners (3.1) and (5.2), alternately using exact inversion or an ILU approximation of the top-left block These results are shown in Figure 6.7.

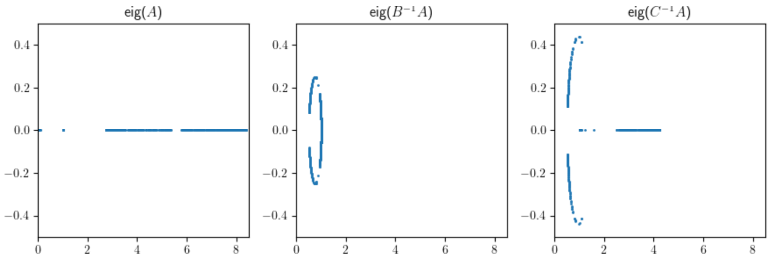

FIG. 6.6. *Eigenvalues of the system matrix A from* (2.38) *(left), and the eigenvalues of the preconditioned system using B from* (3.1) *(middle) and C from* (5.2) *(right).*

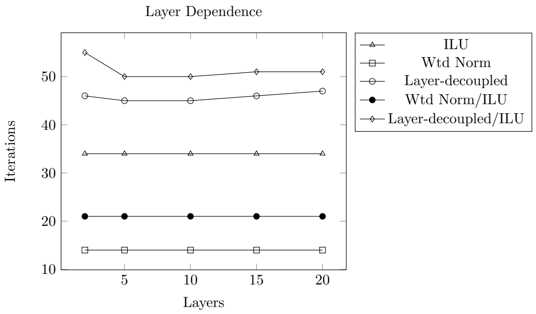

FIG. 6.7. *Iteration count as a function of the number of layers. We see that the performance of our preconditioners seems bounded as we increase the number of layers, a result better than that predicted in Theorem* 5.1.

None of these methods show significant variation as we increase the number of layers. This behavior for (3.1) is not unexpected in light of Theorems 3.1 and 3.2, but is better than one expects given Theorem 5.1.

7. Conclusions and future work.

We have presented a new tide model based on the linearized rotating shallow water equations, but with several layers stratified by density. A mixed finite element discretization similar to that for single-layer models [12] gives rise to a large system of equations, with additional complexity arising from the all-to-all coupling between the layers. We have presented and analyzed weighted-norm preconditioners that are robust with respect to most of the physical and discretization parameters. For typical parameter values, additional approximations such as neglecting interlayer coupling and approximating inverses of matrix blocks with incomplete factorizations may result in highly practical methods.

Future directions for this work would include careful energy-type estimates that sharply describe the physical damping in the system. These would inform a priori estimates like we have previously derived in the single-layer case. Additionally, adapting such energy estimates to the fully discrete case, as well as studying the systems arising from higher-order temporal methods, present further challenges.

REFERENCES

[1] M. S. Alnæs, A. Logg, K. B. Ølgaard, M. E. Rognes, and G. N. Wells, *Unified Form Language: A domain-specific language for weak formulations of partial differential equations*, ACM Trans. Math. Software, 40 (2014), pp. 1–37.

[2] D. N. Arnold, D. Boffi, and R. S. Falk, *Quadrilateral $H(div)$ finite elements*, SIAM J. Numer. Anal., 42 (2005), pp. 2429–2451.

[3] D. N. Arnold, R. S. Falk, and R. Winther, *Preconditioning in $H(div)$ and applications*, Math. Comput., 66 (1997), pp. 957–984, https://doi.org/10.1090/S0025-5718-97-00826-0.

[4] E. Audusse, *A multilayer Saint-Venant model: Derivation and numerical validation*, Discrete Contin. Dyn. Syst. B, 5 (2005), pp. 189–214.

[5] D. Boffi, F. Brezzi, M. Fortin, et al., *Mixed Finite Element Methods and Applications*, Springer Ser. Comput. Math. 44, Springer, Berlin, 2013.

[6] L. Bonaventura, E. D. Fernández-Nieto, J. Garres-Díaz, and G. Narbona-Reina, *Multilayer shallow water models with locally variable number of layers and semi-implicit time discretization*, J. Comput. Phys., 364 (2018), pp. 209–234.

[7] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 3rd ed., Texts Appl. Math. 15, Springer, New York, 2008.

[8] F. Brezzi, D. J. Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[9] R. Comblen, J. Lambrechts, J.-F. Remacle, and V. Legat, *Practical evaluation of five partly discontinuous finite element pairs for the non-conservative shallow water equations*, Internat. J. Numer. Methods Fluids, 63 (2010), pp. 701–724.

[10] C. Cotter and D. Ham, *Numerical wave propagation for the triangular P1DG-P 2 finite element pair*, J. Comput. Phys., 230 (2011), pp. 2806–2820, https://doi.org/DOI:10.1016/j.jcp.2010.12.024.

[11] C. J. Cotter, P. J. Graber, and R. C. Kirby, *Mixed finite elements for global tide models with nonlinear damping*, Numer. Math., 140 (2018), pp. 963–991.

[12] C. J. Cotter and R. C. Kirby, *Mixed finite elements for global tide models*, Numer. Math., 133 (2016), pp. 255–277.

[13] C. J. Cotter and J. Shipton, *Mixed finite elements for numerical weather prediction*, J. Comput. Phys., 231 (2012), pp. 7076–7091.

[14] J. Crum, C. Cheng, D. A. Ham, L. Mitchell, R. C. Kirby, J. A. Levine, and A. Gillette, *Bringing trimmed serendipity methods to computational practice in Firedrake*, ASM Trans. Math. Software, 48 (2022), pp. 1–19.

[15] G. D. Egbert and R. D. Ray, *Significant dissipation of tidal energy in the deep ocean inferred from satellite altimeter data*, Nature, 405 (2000), pp. 775–778.

[16] P. E. Farrell, R. C. Kirby, and J. Marchena-Menendez, *Irksome: Automating Runge-Kutta time-stepping for finite element methods*, ACM Trans. Math. Software, 47 (2021), pp. 1–26.

[17] C. Garrett and E. Kunze, *Internal tide generation in the deep ocean*, Annu. Rev. Fluid Mech., 39 (2007), pp. 57–87.

[18] A. Gillette, T. Kloefkorn, and V. Sanders, *Computational serendipity and tensor product finite element differential forms*, SMAI J. Comput. Math., 5 (2019), pp. 1–21.

[19] S. D. Griffiths and R. H. Grimshaw, *Internal tide generation at the continental shelf modeled using a modal decomposition: Two-dimensional results*, J. Phys. Oceanogr., 37 (2007), pp. 428–451.

[20] R. C. Kirby, *From functional analysis to iterative methods*, SIAM Rev., 52 (2010), pp. 269–293.

[21] R. C. Kirby and T. Kernell, *Preconditioning mixed finite elements for tide models*, Comput. Math. Appl., 82 (2021), pp. 212–227.

[22] R. C. Kirby and L. Mitchell, *Solver composition across the PDE/linear algebra barrier*, SIAM J. Sci. Comput., 40 (2018), pp. C76–C98, https://doi.org/10.1137/17M1133208.

[23] D. Y. Le Roux, V. Rostand, and B. Pouliot, *Analysis of numerically induced oscillations in 2D finite-element shallow-water models part I: Inertia-gravity waves*, SIAM J. Sci. Comput., 29 (2007), pp. 331–360.

[24] D. Y. Le Roux, *Dispersion relation analysis of the $P_1^{NC} - P_1$ finite-element pair in shallow-water models*, SIAM J. Sci. Comput., 27 (2005), pp. 394–414.

[25] D. Y. Le Roux and B. Pouliot, *Analysis of numerically induced oscillations in two-dimensional finite-element shallow-water models part II: Free planetary waves*, SIAM J. Sci. Comput., 30 (2008), pp. 1971–1991.

[26] J. Le Sommer, S. Medvedev, R. Plougonven, and V. Zeitlin, *Singularity formation during relaxation of jets and fronts toward the state of geostrophic equilibrium*, Commun. Nonlinear Sci. Numer. Simul., 8 (2003), pp. 415–442.

[27] K. T. Mandli, *Finite Volume Methods for the Multilayer Shallow Water Equations with Applications to Storm Surges*, Ph.D. thesis, University of Washington, Seattle, WA, 2011.

[28] K.-A. Mardal and R. Winther, *Preconditioning discretizations of systems of partial differential equations*, Numer. Linear Algebra Appl., 18 (2011), pp. 1–40.

[29] W. Munk and C. Wunsch, *Abyssal recipes* II*: Energetics of tidal and wind mixing*, Deep-Sea Res. Part I, 45 (1998), pp. 1977–2010.

[30] F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. T. McRae, G.-T. Bercea, G. R. Markall, and P. H. J. Kelly, *Firedrake: Automating the finite element method by composing abstractions*, ACM Trans. Math Software, 43 (2016), 24, https://doi.org/10.1145/2998441.

[31] P. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Rome, 1975, Lecture Notes in Math. 606, Springer, Berlin, 1977, pp. 292–315.

[32] V. Rostand and D. Le Roux, *Raviart-Thomas and Brezzi-Douglas-Marini finite-element approximations of the shallow-water equations*, Internat. J. Numer. Methods Fluids, 57 (2008), pp. 951–976.

[33] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2003.

[34] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[35] L. St. Laurent and C. Garrett, *The role of internal tides in mixing the deep ocean*, J. Phys. Oceanogr., 32 (2002), pp. 2882–2899.

[36] G. Strang, *Introduction to Linear Algebra*, Vol. 3, Wellesley-Cambridge Press, Wellesley, MA, 1993.

[37] H. Weller, T. Ringler, M. Piggott, and N. Wood, *Challenges facing adaptive mesh modeling of the atmosphere and ocean*, Bull. Amer. Meteorol. Soc., 91 (2010), pp. 105–108.