Semi-Implicit Time Differencing

Scott R. Fulton

Revised June 1, 2004

Abstract

The basic ideas of semi-implicit time differencing are reviewed, including analytical results not widely referenced in the meteorological literature. It is shown that solving the resulting implicit equations by simple fixed-point iteration requires a relatively small time step to guarantee convergence, thus defeating the purpose of the semi-implicit method. In contrast, classical iteration methods (e.g., Jacobi, Gauss-Seidel) converge regardless of the time step if the implicit problem is elliptic. These ideas are illustrated in detail for the shallow-water equations in one dimension and sketched for the two-dimensional case.

A detailed analysis of absolute stability is combined with previously derived conditions on the order of accuracy to identify promising schemes for the time integration of meteorological models. All appropriate cases of combined linear multistep methods of orders up to three (explicit) and two (implicit) are considered. In particular, a stable and promising semi-implicit formulation of the third-order Adams-Bashforth method is proposed.

Technical Report No. 2002-01
Department of Mathematics and Computer Science
Clarkson University, Potsdam, NY 13699–5815

This work was supported by

the U. S. Department of Energy
Subcontract to DOE Cooperative Agreement DE-FC02-01ER63163

and by

the Office of Naval Research
Marine Meteorology and Atmospheric Effects Program
Grants N00014-98-1-0103 and N00014-98-1-0368

1 Introduction

Semi-implicit time differencing has been widely used in meteorological models starting with Kwizak and Robert [12] and Robert et al. [13]. The basic idea is to discretize a time-dependent system of equations using an implicit scheme for some terms and an explicit scheme for the remaining terms. The goal is to split up the terms in such a way that the largest stable time step for the semi-implicit discretization is significantly larger than for a corresponding explicit discretization, thus reducing the computer time required to solve the equations.

To achieve this goal without sacrificing accuracy, three issues must be addressed:

1. all terms required for the fastest solution components should be treated implicitly,
2. the energy in those fast components should be insignificant, and
3. the resulting implicit system of equations should be relatively easy to solve.

Points 1 and 3 should be obvious. Point 2 is not always clearly stated, but it is necessary: solution modes treated implicitly—while stable—will still be distorted by the time discretization, and thus an accurate solution can be obtained only if there is little energy in these modes. In the classical case of "stiff" equations (e.g., the solution contains some modes which decay very quickly), the fast decaying modes will inevitably contain little energy and thus (usually) can be treated implicitly. In contrast, the usual meteorological situation has waves of two time scales (e.g., Rossby waves and gravity waves), and accurate results can be obtained by semi-implicit time differencing only when the fast waves (treated implicitly) contain little energy.

This paper reviews the basic ideas of semi-implicit time differencing and its application to meteorological models. Section 2 introduces a general setting for semi-implicit schemes, gives examples of schemes used in practice, and surveys some important convergence results. Section 3 compares methods for solving the resulting implicit equations. Applications of these ideas to the shallow water model in one and two dimensions are discussed in section 4. A general formulation of combined linear multistep methods is presented in section 5, including the analysis of convergence and order of accuracy. Section 6 details the analysis of absolute stability for semi-implicit schemes (to third-order) and identifies promising high-order methods. Our conclusions are summarized in section 7.

var / dim	equation(s)	space continuous	space discrete
1 / 1	$u_t + u_x = 0$	$\psi = u \in C^1(\Omega)$	$\psi = \mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix} \in \mathbb{R}^N$
2 / 1	$u_t + gh_x = 0$ $h_t + \bar{h}u_x = 0$	$\psi = \begin{bmatrix} u \\ h \end{bmatrix} \in \left[C^1(\Omega)\right]^2$	$\psi = \begin{bmatrix} \mathbf{u} \\ \mathbf{h} \end{bmatrix} \in \mathbb{R}^{2N}$
3 / 2	$u_t - fv + gh_x = 0$ $v_t + fu + gh_y = 0$ $h_t + \bar{h}(u_x + v_y) = 0$	$\psi = \begin{bmatrix} u \\ v \\ h \end{bmatrix} \in \left[C^1(\Omega)\right]^3$	$\psi = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{h} \end{bmatrix} \in \mathbb{R}^{3N^2}$

Table 1: Interpretation of the variable $\psi(t)$ and the space S in which it lies for various equations of the form (2.1). Here Ω denotes the spatial domain of the problem and partial derivatives are denoted by subscripts. Space-discrete forms are illustrated assuming N gridpoint values (or spectral coefficients) in each coordinate direction for simplicity.

2 Overview

For a reasonably general setting for semi-implicit schemes,[1] consider a system of equations of the form

$$\psi'(t) = \mathcal{A}(\psi(t)) + \mathcal{B}(\psi(t)). \tag{2.1}$$

Here $\psi(t)$ represents the solution, which at each time $t \geq 0$ lies in some vector space S, and the operators $\mathcal{A}$ and $\mathcal{B}$ defined on S represent terms to be treated implicitly and explicitly, respectively. We intend (2.1) to represent a system of time-dependent partial differential equations, either before or after discretization in space (by any method). Note that we assume the space and time discretizations can be separated completely, so our approach here is that commonly referred to as the "method of lines." The vector space S in which $\psi(t)$ lies depends on the number of variables, whether the equations are continuous or discrete in space, the order of the space derivative(s), etc. Some examples are given in Table 1; the shallow-water equations in one and two dimensions (rows two and three of this Table) are treated in detail in sections 4.2 and 4.3. The problem (2.1) is an autonomous system; it could be generalized by making it non-autonomous (i.e., including explicit t dependence in $\mathcal{A}$ and $\mathcal{B}$), including "algebraic" (i.e., diagnostic) equations to form a differential-algebraic equation, and/or including more terms on the right. However, this form is sufficient to illustrate the ideas to be considered here.

[1] Such methods are also referred to in the literature as "implicit-explicit methods" [2], "additive methods" [4, 5], and "combined methods" [11].

We consider discretizations of (2.1) which replace the solution $\psi(t)$ with an approximation $\psi^n \approx \psi(t_n)$ at times $t_0 < t_1 < \cdots < t_n < \cdots$. For simplicity here we will take $t_n = n\Delta t$, where the time step Δt is constant, but variable time steps could also be used. A *semi-implicit* discretization treats $\mathcal{A}(\psi)$ implicitly and $\mathcal{B}(\psi)$ explicitly, i.e., it reduces to an explicit method if $\mathcal{A}(\psi) \equiv 0$ and to an implicit method if $\mathcal{B}(\psi) \equiv 0$. For example, combining the backward (implicit Euler) and forward (explicit Euler) schemes yields the discretization

$$\frac{\psi^{n+1} - \psi^n}{\Delta t} = \mathcal{A}(\psi^{n+1}) + \mathcal{B}(\psi^n). \tag{2.2}$$

The scheme most often used in meteorological models (e.g., [6, 15, 17]) combines the trapezoidal (implicit) and leapfrog (explicit) schemes, yielding the discretization

$$\frac{\psi^{n+1} - \psi^{n-1}}{2\Delta t} = \frac{\mathcal{A}(\psi^{n-1}) + \mathcal{A}(\psi^{n+1})}{2} + \mathcal{B}(\psi^n). \tag{2.3}$$

It is possible to use more general linear multistep (LM) methods. For example, combining a general weighted average for the implicit term and the second-order Adams-Bashforth (explicit) scheme (AB2) for the explicit term yields the discretization used in [16], i.e.,

$$\frac{\psi^{n+1} - \psi^n}{\Delta t} = (1-\theta)\mathcal{A}(\psi^n) + \theta\mathcal{A}(\psi^{n+1}) + \frac{3}{2}\mathcal{B}(\psi^n) - \frac{1}{2}\mathcal{B}(\psi^{n-1}). \tag{2.4}$$

Here the values $\theta = \frac{1}{2}$ and $\theta = 1$ give the trapezoidal and backward implicit schemes, respectively.[2] It is also possible to use Runge-Kutta (RK) methods, e.g., combining the backward (implicit Euler) and classical fourth-order Runge-Kutta (explicit) schemes (see details in [8]). In general, most implicit and explicit schemes may be combined, with some natural restrictions (see [11] for details).

The first question to address with any discretization of differential equations is that of *convergence*: if we fix t and use n time steps to compute $\psi^n \approx \psi(t)$ with $\Delta t = t/n$, does ψ^n converge to $\psi(t)$ as $n \to \infty$, i.e., as $\Delta t \to 0$? Furthermore, if the method converges, what is the order of convergence? These fundamental questions—apparently not addressed in the meteorological literature—have been answered for quite general semi-implicit methods [2, 4, 5, 11]. The following basic results have been established for combinations of two LM methods or two RK methods:

- If both methods are convergent separately, their combination is convergent.
- With LM methods the order is the minimum of the orders of the two methods.
- With RK methods the order may be *less than* the minimum of the orders of the two methods.
- If the implicit method has the lower order, then the leading term(s) in the truncation error are proportional to derivatives of $\mathcal{A}(\psi)$ and do not involve $\mathcal{B}(\psi)$.

[2] In the notation of [16], $\theta = (1+\epsilon)/2$, where ϵ is an "uncentering" parameter.

This last point is significant, since it allows using a lower-order implicit scheme when the terms treated implicitly contain little energy. For example, when dissipative terms are treated implicitly, the backward Euler scheme (only first-order) may be a better choice than the trapezoidal scheme (which is second-order), since solution modes which should decay quickly do so in the former but are nearly constant in the latter.

We consider the questions of convergence and order of accuracy in more detail below (section 5), focusing on the case of combined linear multistep methods (such as the examples given above). We also treat the more complicated question of absolute stability (section 6). However, we first consider the practical question of how to solve the resulting implicit system of equations (section 3), and illustrate these ideas by applying them to the shallow-water equations (section 4).

3 Solution of the Implicit Equations

Using a semi-implicit discretization such as (2.2) or (2.3) requires solving an implicit problem of the form

$$\psi - \gamma \Delta t \mathcal{A}(\psi) = f. \tag{3.1}$$

Here ψ denotes the solution sought at time level $n+1$ (with the superscript dropped for simplicity), f denotes known terms (computed explicitly from values at time level n and previous levels), and γ is a constant which depends on the implicit scheme used, e.g., $\gamma = 1$ with for the backward Euler scheme, $\gamma = \frac{1}{2}$ for the trapezoidal scheme, and $\gamma = \theta$ for the more general weighting in (2.4). In spectral models (e.g., [16]) this equation is expressed in spectral space where it is often diagonal and thus trivial to solve. However, with other discretizations the question of how to solve (3.1) is critical: we need a method which works and does not cost too much (to avoid compromising overall efficiency). In the following subsections we outline the solution of (3.1) by simple fixed-point iteration (which is usually unsatisfactory) and by other iterative methods (which usually work).

3.1 Fixed-point iteration

From (3.1) we see that the solution ψ (if it exists) is a *fixed point* of the function

$$g(\psi) := f + \gamma \Delta t \mathcal{A}(\psi), \tag{3.2}$$

i.e., that $g(\psi) = \psi$. This suggests using simple *fixed-point iteration*[3] as follows: starting with an initial guess $\psi^{(0)}$ for ψ, generate the sequence $\psi^{(k)}$, $k = 0, 1, \dots$ via

$$\psi^{(k+1)} = g(\psi^{(k)}) = f + \gamma \Delta t \mathcal{A}(\psi^{(k)}). \tag{3.3}$$

Note that here the superscript in parentheses indexes successive approximations to the solution $\psi = \psi^{n+1}$. This method is attractive, since it avoids the need to solve any implicit equations.

For the iteration (3.3) to be useful it must converge, i.e., we must have $\|\psi - \psi^{(k)}\| \to 0$ as $k \to \infty$ in some appropriate norm $\| \cdot \|$. From the general theory of fixed-point iteration (e.g., [18]), it can be shown that a unique solution exists and the iteration converges to it provided g is a contraction on a compact subset of the solution space S. In practice, this requires that

$$\|g'(\psi)\| < 1 \tag{3.4}$$

where $g'(\psi)$ is the functional derivative at the fixed point (i.e., solution) ψ and the norm is the corresponding operator norm. Indeed, the convergence is linear and $\|g'(\psi)\|$ (assumed

[3] Sometimes referred to as *functional iteration*

to be nonzero) is the asymptotic convergence factor, i.e., the factor by which the error is reduced per iteration. Substituting from (3.2) leads to the condition $\gamma\Delta t\|\mathcal{A}'(\psi)\| < 1$, i.e.,

$$\Delta t < \frac{1}{\gamma\|\mathcal{A}'(\psi)\|} \tag{3.5}$$

as a necessary condition for convergence (in general, this condition is also sufficient). Note that this condition plays the same practical role as the stability (CFL) condition which the semi-implicit scheme was supposed to avoid, i.e., it limits the time step that can be taken. Thus, attempting to avoid the difficulty of solving the implicit equation (3.1) by using simple fixed-point iteration defeats the purpose of semi-implicit time differencing.

3.2 Other iterative schemes

If simple fixed-point iteration is not satisfactory, how can the implicit equation (3.1) be solved? In general, direct solution is impractical due to the size of the problem (especially in more than one space dimension), and iterative methods are appropriate. The key here is that conventional iterative methods *couple the time and space discretizations*, thereby avoiding the limitation (3.5) on the time step. For concreteness, we consider the case where the equations have been discretized in space so $\psi(t) = \mathbf{v} \in \mathbb{R}^N$ for some N and the implicit equation (3.1) to be solved takes the form

$$\mathbf{v} - \gamma\Delta t\mathbf{A}(\mathbf{v}) = \mathbf{f}, \tag{3.6}$$

where we have used boldface symbols for vectors in $\mathbb{R}^N$.

If the term treated implicitly is linear, i.e.,

$$\mathbf{A}(\mathbf{v}) = A\mathbf{v} \tag{3.7}$$

where $A \in \mathbb{R}^{N\times N}$, then (3.6) becomes

$$\mathbf{v} - \gamma\Delta tA\mathbf{v} = \mathbf{f}. \tag{3.8}$$

The classical Jacobi and Gauss-Seidel iterative methods are based on the matrix splitting $A = D + L + U$, where D, L, and U are matrices consisting of the diagonal, lower triangle, and upper triangle of A, respectively. For the Jacobi method we write (3.8) as

$$(I - \gamma\Delta tD)\mathbf{v} = \mathbf{f} + \gamma\Delta t(L + U)\mathbf{v}, \tag{3.9}$$

leading to the iteration

$$\mathbf{v}^{(k+1)} = (I - \gamma\Delta tD)^{-1}\left[\mathbf{f} + \gamma\Delta t(L + U)\mathbf{v}^{(k)}\right]. \tag{3.10}$$

Similarly, for the Gauss-Seidel method we write (3.8) as

$$\left[I - \gamma\Delta t(D + L)\right]\mathbf{v} = \mathbf{f} + \gamma\Delta tU\mathbf{v}, \tag{3.11}$$

leading to the iteration

$$\mathbf{v}^{(k+1)} = [I - \gamma\Delta t(D+L)]^{-1}\left(\mathbf{f} + \gamma\Delta t U\mathbf{v}^{(k)}\right). \tag{3.12}$$

Other iterative methods such as SOR, ADI, and multigrid cycling can be formulated similarly.

Each such iteration is a fixed-point iteration of the form

$$\mathbf{v}^{(k+1)} = \mathbf{g}\left(\mathbf{v}^{(k)}\right) := M\mathbf{v}^{(k)} + \mathbf{c}, \tag{3.13}$$

where M is an *iteration matrix* and $\mathbf{c}$ is a constant vector. For the Jacobi method we have

$$M_J := \gamma\Delta t(I - \gamma\Delta t D)^{-1}(L+U), \qquad \mathbf{c}_J := (I - \gamma\Delta t D)^{-1}\mathbf{f}, \tag{3.14}$$

and for the Gauss-Seidel method we have

$$M_G := \gamma\Delta t\,[I - \gamma\Delta t(D+L)]^{-1}\,U. \qquad \mathbf{c}_G := [I - \gamma\Delta t(D+L)]^{-1}\,\mathbf{f}. \tag{3.15}$$

Since the error $\mathbf{v} - \mathbf{v}^{(k)}$ satisfies

$$\mathbf{v} - \mathbf{v}^{(k)} = M\left(\mathbf{v} - \mathbf{v}^{(k-1)}\right) = M^k\mathbf{v}^{(0)}, \tag{3.16}$$

the iteration converges (for any initial approximation $\mathbf{v}^{(0)}$) if and only if $\rho(M) < 1$, where ρ denotes the spectral radius; this condition implies that $\|M\| < 1$ in some operator matrix norm. Indeed, since the iteration function $\mathbf{g}$ has derivative $\mathbf{g}'(\mathbf{v}) = M$, this condition for convergence matches precisely with (3.4). However, unlike the simple fixed-point iteration (3.3), these iteration functions couple the space and time discretizations: part of the space discretization term $\mathbf{A}(\mathbf{v})$ is combined directly with the term $\mathbf{v}$ coming from the time derivative [cf. (3.9) and (3.11)].

For nonlinear problems, one can use Newton's method. Specifically, a solution of (3.6) is a root of the equation $\mathbf{F}(\mathbf{v}) = \mathbf{v} - \gamma\Delta t\mathbf{A}(\mathbf{v}) - \mathbf{f} = 0$, for which Newton's method takes the form:

```
For $k = 0, 1, \ldots$ do:
    solve $\mathbf{F}'(\mathbf{v}^{(k)})\delta\mathbf{v}^{(k)} = -\mathbf{F}(\mathbf{v}^{(k)})$ for $\delta\mathbf{v}^{(k)}$
    set $\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \delta\mathbf{v}^{(k)}$
```

If needed, a similar quasi-Newton or damped Newton method could be used. The main point here is that the linear system to be solved in each Newton iteration involves the Jacobian matrix

$$\mathbf{F}'(\mathbf{v}^{(k)}) = I - \gamma\Delta t A^{(k)} \tag{3.17}$$

where $A^{(k)} = \mathbf{A}'(\mathbf{v}^{(k)})$; since these systems have precisely the same form as (3.8), they could be solved in a similar manner.

With general iterative methods such as those outlined above, we cannot find conditions like (3.5) which guarantee convergence without knowing the details of the term $\mathbf{A}(\mathbf{v})$, i.e., the equations and spatial discretization. However, in general one finds that whenever the implicit problem (3.1) is linear and elliptic and is discretized appropriately to obtain (3.6), the classical iterative methods all converge *for any time step* Δt, thus avoiding the limitation (3.5) on the time step. An example is given in the next section.

4 Application to the Shallow Water Equations

The shallow-water system includes advective, Coriolis, and pressure gradient terms, and thus serves as a prototype for general primitive-equation models. We can write the momentum and mass continuity equations in the form

$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} + f\mathbf{k} \times \mathbf{v} + g\nabla h &= \mathbf{F}, & (4.1) \\ \frac{\partial h}{\partial t} + \nabla \cdot (h\mathbf{v}) &= Q, & (4.2) \end{aligned}$$

where $\mathbf{v}$ is the (horizontal) vector velocity,[4] h is the free surface height, f is the Coriolis parameter, g is the gravitational constant, $\mathbf{k}$ is the vertical unit vector, and the operator ∇ is restricted to the horizontal. The source terms $\mathbf{F}$ and Q for momentum and mass are regarded as known functions of t (and possibly of $\mathbf{v}$ and h).

In this section we use the shallow-water equations to illustrate the concepts and conclusions of the previous sections. First, we use a linear analysis to identify a reasonable splitting of the terms. Next, we describe a simple discretization for the linear case in one dimension and verify the arguments of section 3, namely, that for solving the implicit problem, fixed-point iteration does not work but other standard iterative methods do. Finally, we show how to formulate and solve the implicit problem for the nonlinear case in two dimensions.

4.1 Linear analysis

To justify a proper splitting of terms for semi-implicit time differencing, we use the following linear analysis.[5] Working on a f-plane (Cartesian coordinates with f constant) with velocity components u and v in the x and y directions and no forcing, we can linearize (4.1)–(4.2) about a basic state with velocity $(\bar{u}, \bar{v})$ and height $\bar{h} > 0$ (both constant) to obtain

$$\begin{aligned} u_t + \bar{u}u_x + \bar{v}u_y - fv + gh_x &= 0, & (4.3) \\ v_t + \bar{u}v_x + \bar{v}v_y + fu + gh_y &= 0, & (4.4) \\ h_t + \bar{u}h_x + \bar{v}h_y + \bar{h}(u_x + v_y) &= 0, & (4.5) \end{aligned}$$

where subscripts denote partial derivatives and h now denotes the *deviation* of the free surface height from the reference height $\bar{h}$. Scaling h by g/c, where $c = \sqrt{g\bar{h}}$ is the phase speed of a pure gravity wave, we can write (4.3)–(4.5) in the symmetrized matrix form

$$\frac{\partial \psi}{\partial t} = \mathcal{L}\psi, \qquad (4.6)$$

[4] Here, boldface symbols denote vectors in physical space rather than vectors in $\mathbb{R}^N$.

[5] A linear analysis based on the vorticity-divergence form yields the same conclusions.

where $\psi = [u, v, gh/c]^T$ and

$$\mathcal{L} = \begin{bmatrix} -\bar{u}\partial_x - \bar{v}\partial_y & f & -c\partial_x \\ -f & -\bar{u}\partial_x - \bar{v}\partial_y & -c\partial_y \\ -c\partial_x & -c\partial_y & -\bar{u}\partial_x - \bar{v}\partial_y \end{bmatrix}. \tag{4.7}$$

Assuming wave-form solutions, i.e., $\psi(x, y, t) = \hat{\psi}(t)e^{i(kx+ly)}$, leads to

$$\frac{d\hat{\psi}}{dt} = L\hat{\psi}, \tag{4.8}$$

where

$$L = \begin{bmatrix} -i\bar{\omega} & f & -ikc \\ -f & -i\bar{\omega} & -ilc \\ -ikc & -ilc & -i\bar{\omega} \end{bmatrix} \tag{4.9}$$

with $\bar{\omega} := k\bar{u} + l\bar{v}$ (which we assume is nonnegative). Since L is skew-Hermitian, its eigenvalues are pure imaginary. Writing this eigenvalue problem in the form $L\Psi_j = -i\omega_j\Psi_j$ for $j = 1,\ 2,\ 3$ (with ω_j real) allows us to express the general solution for ψ as a linear combination of solutions of the form

$$\psi(x, y, t) = \Psi_j e^{i(kx+ly-\omega_j t)}. \tag{4.10}$$

Direct calculation yields the eigenvalues and corresponding eigenvectors

$$\begin{aligned} &\text{Slow (geostrophic) modes:} \quad && \omega_1 = \bar{\omega}, && \Psi_1 = \frac{1}{\nu}\begin{bmatrix} -ilc \\ ikc \\ f \end{bmatrix}, \\ &\text{Fast (gravity-inertia) modes:} \quad && \omega_{2,3} = \bar{\omega} \pm \nu, && \Psi_{2,3} = \frac{1}{\nu\bar{k}\sqrt{2}}\begin{bmatrix} \pm\nu k + ilf \\ \pm\nu l - ikf \\ c\bar{k}^2 \end{bmatrix}, \end{aligned} \tag{4.11}$$

where $\nu := (f^2 + \omega_g^2)^{1/2}$, $\omega_g = \bar{k}c$, and $\bar{k} := (k^2 + l^2)^{1/2}$. Here the eigenvectors Ψ_j have been normalized so they are orthonormal in the standard inner product $\langle\Psi, \Phi\rangle = \Psi^*\Phi$ (where the asterisk denotes the conjugate transpose).

To determine an appropriate splitting of terms for semi-implicit time differencing, we write (4.8) in the form

$$\frac{d\hat{\psi}}{dt} = A\hat{\psi} + B\hat{\psi} + C\hat{\psi}, \tag{4.12}$$

where

$$A = \begin{bmatrix} 0 & 0 & -ikc \\ 0 & 0 & -ilc \\ -ikc & -ilc & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -i\bar{\omega} & 0 & 0 \\ 0 & -i\bar{\omega} & 0 \\ 0 & 0 & -i\bar{\omega} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & f & 0 \\ -f & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{4.13}$$

The matrices A, B, and C correspond to the gravity-wave, advective, and Coriolis terms, respectively. The corresponding spectral radii $\rho(A) = \omega_g$, $\rho(B) = \bar{\omega}$, and $\rho(C) = f$ typically satisfy

$$\omega_g \gg \bar{\omega} \gg f, \tag{4.14}$$

corresponding to a separation of time scales. In particular, the ratio $\rho(A)/\rho(B) = \omega_g/\bar{\omega} = c/\bar{u}$ is the reciprocal of the Froude number; this is typically about 10. For example, using the values $c \sim 300\,\mathrm{ms}^{-1}$, $\bar{u} \sim 30\,\mathrm{ms}^{-1}$, $f \sim 10^{-5}\,\mathrm{s}^{-1}$, and $\bar{k} \sim 2\pi/(200\,\mathrm{km})$ leads to $\omega_g/\bar{\omega} \sim 10$ and $\bar{\omega}/f \sim 100$. Thus, we can think of the terms A, B, and C in (4.12) as the "fast", "slower", and "slowest" terms, respectively.

More precisely, we can compute the size of each of the terms on the right in (4.12) for each of the modes Ψ_j ($j = 1, 2, 3$). Using the standard norm $\|\Psi\| = \sqrt{\langle \Psi, \Psi \rangle}$, for the slow (geostrophic) modes ($j = 1$) we obtain

$$\|A\Psi_1\| = \frac{f\omega_g}{\nu} \approx f, \qquad \|B\Psi_1\| = \bar{\omega}, \qquad \|C\Psi_1\| = \frac{f\omega_g}{\nu} \approx f, \tag{4.15}$$

where we have used $f \ll \omega_g$ and thus $\omega_g \approx \nu$. Similarly, for the fast (gravity-inertia) modes ($j = 2, 3$) we obtain

$$\|A\Psi_{2,3}\| = \frac{\omega_g}{\nu}\left(\frac{\omega_g^2 + \nu^2}{2}\right)^{1/2} \approx \nu, \quad \|B\Psi_{2,3}\| = \bar{\omega}, \quad \|C\Psi_{2,3}\| = \frac{f}{\nu}\left(\frac{f^2 + \nu^2}{2}\right)^{1/2} \approx \frac{f}{\sqrt{2}}. \tag{4.16}$$

Using the scales assumed above we thus find that the Coriolis terms (C) have (approximate) size f and the advection terms (B) have size $1000f$ for both the slow and fast modes, while the gravity-wave terms (A) have (approximate) size f for the slow modes but $10f$ for the fast modes. Thus, the terms we have identified as the "gravity-wave" terms indeed are most important for the fastest components of the solution; it will be reasonable to treat these terms implicitly in situations where the fast modes contain little energy. On the other hand, the advective terms should be treated explicitly, since they are largest and their time scale is an order of magnitude slower.

One question remains: should we treat the Coriolis terms implicitly or explicitly? On one hand, their time scale is much slower, so there is no apparent need to treat them implicitly. On the other hand, since they are linear, including them in the problem to be solved implicitly might not be difficult and perhaps have some advantage (e.g., smaller truncation error). Here, the above analysis suggests a preference for the latter as follows. If we treat the Coriolis terms implicitly, this means splitting the operator L in (4.8) into $A + C$ and B. Since $B = -i\bar{\omega}I$, it commutes with $A + C$, so these operators are simultaneously orthogonally diagonalizable [14]. This means that the absolute stability analysis of section 6 applies directly (at least in this linear problem), giving some confidence that methods identified as good by analysis will in fact work properly. On the other hand, if we treat the Coriolis terms explicitly, this means splitting the operator L in (4.8) into A and $B + C$. Since these two operators do not commute,[6] they are not simultaneously orthogonally diagonalizable, and the absolute stability analysis is somewhat suspect. In this sense it would be natural to treat the Coriolis terms implicitly. In practice, since the contribution from these terms is relatively small, there may be little difference.

[6] Since A, B, and C are individually skew-Hermitian, $A(B + C) = [(B + C)A]^* \neq (B + C)A$.

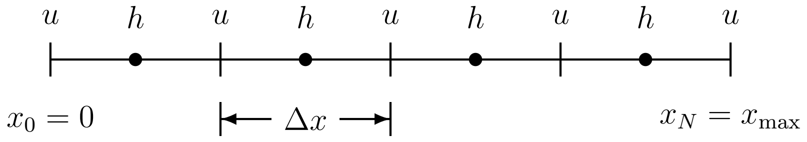

Figure 1: Variable staggering for one-dimensional shallow-water model.

4.2 One-dimensional shallow-water equations

To illustrate methods for solving the implicit problem in a simple context, consider the linearized shallow-water equations in one space dimension x, i.e., (4.3)–(4.5) with all y-derivatives set to zero. Dropping the Coriolis terms for simplicity and setting $\bar{u} = 0$ so we can focus entirely on the implicit (gravity-wave) terms, this model may be written as

$$\begin{aligned} u_t + g h_x &= 0, \qquad (4.17)\\ h_t + \bar{h} u_x &= 0. \qquad (4.18) \end{aligned}$$

4.2.1 Discretization

Using N grid intervals on the domain $0 \le x \le x_{\max}$ with $x_j = j\Delta x = j x_{\max}/N$ for $j = 0, \ldots, N$ and variables staggered as shown in Fig. 1, the problem can be discretized using centered finite differences as

$$\begin{aligned} \frac{du_j}{dt} + g\frac{h_{j+\frac{1}{2}} - h_{j-\frac{1}{2}}}{\Delta x} &= 0, \qquad j = 1, \ldots, N-1, \qquad (4.19)\\ \frac{dh_{j-\frac{1}{2}}}{dt} + \bar{h}\frac{u_j - u_{j-1}}{\Delta x} &= 0, \qquad j = 1, \ldots, N, \qquad (4.20)\\ & \qquad (4.21) \end{aligned}$$

where $u_j(t) \approx u(x_j, t)$ and $h_{j-\frac{1}{2}}(t) \approx h(x_{j-\frac{1}{2}}, t)$. The wall boundary conditions $u_0 = u_N = 0$ complete the discrete system.

Writing the unknowns in vector form as $\mathbf{u} = [u_1, \ldots, u_{N-1}]^T$ and $\mathbf{h} = [h_{\frac{1}{2}}, \ldots, h_{N-\frac{1}{2}}]^T$, we can write the discrete equations in matrix form as

$$\begin{aligned} \frac{d\mathbf{u}}{dt} - \frac{g}{\Delta x} D\mathbf{h} &= 0, \qquad (4.22)\\ \frac{d\mathbf{h}}{dt} + \frac{\bar{h}}{\Delta x} D^T\mathbf{u} &= 0, \qquad (4.23) \end{aligned}$$

where D is the $(N-1)\times N$ matrix

$$D = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix} \tag{4.24}$$

(this matrix has no relation to the D used in section 3.2). Then defining

$$\mathbf{v} = \begin{bmatrix} \mathbf{u} \\ \mathbf{h} \end{bmatrix}, \qquad A = \frac{1}{\Delta x}\begin{bmatrix} 0 & gD \\ -\bar{h}D^T & 0 \end{bmatrix} \tag{4.25}$$

we can write the space-discretized equations as

$$\frac{d\mathbf{v}}{dt} = A\mathbf{v}. \tag{4.26}$$

This corresponds to the generic problem (2.1) with $\mathbf{v}$ playing the role of ψ and $\mathcal{A}(\mathbf{v}) = A\mathbf{v}$; there are no terms to treat explicitly, since we have neglected the advection and Coriolis terms. Consequently, the implicit problem (3.1) to be solved with (semi-)implicit time differencing has the form

$$\mathbf{v} - \gamma\Delta t A\mathbf{v} = \mathbf{f} \tag{4.27}$$

[cf. (3.8)], where γ is a constant (e.g., $\gamma = \frac{1}{2}$ for the Trapezoidal method) and right-hand side $\mathbf{f}$ is known.

4.2.2 Fixed-point iteration

The simple fixed-point iteration (3.3) here takes the form

$$\mathbf{v}^{(k+1)} = \mathbf{g}(\mathbf{v}^{(k)}) := \mathbf{f} + \gamma\Delta t A\mathbf{v}^{(k)}, \tag{4.28}$$

which can be expanded as

$$\begin{aligned} \mathbf{u}^{(k+1)} &= \mathbf{f}_u + \frac{g\gamma\Delta t}{\Delta x} D\mathbf{h}^{(k)}, \\ \mathbf{h}^{(k+1)} &= \mathbf{f}_h - \frac{\bar{h}\gamma\Delta t}{\Delta x} D^T\mathbf{u}^{(k)}, \end{aligned} \tag{4.29}$$

where $\mathbf{f}_u$ and $\mathbf{f}_h$ are the corresponding parts of the known right-hand side $\mathbf{f}$. The convergence factor is $\rho(\mathbf{g}'(\mathbf{v})) = \gamma\Delta t\rho(A)$. To find $\rho(A)$ we write $A\mathbf{v} = -i\omega\mathbf{v}$ (where we have anticipated that the eigenvalues will be pure imaginary) and expand this to obtain

$$\frac{g}{\Delta x} D\mathbf{h} = -i\omega\mathbf{u}, \tag{4.30}$$

$$-\frac{\bar{h}}{\Delta x} D^T\mathbf{u} = -i\omega\mathbf{h}. \tag{4.31}$$

Eliminating $\mathbf{u}$ gives

$$D^T D\mathbf{h} = \left(\frac{\omega\Delta x}{c}\right)^2 \mathbf{h} \tag{4.32}$$

where $c = \sqrt{g\bar{h}}$ is the gravity wave speed. Since

$$D^T D = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{N\times N}, \tag{4.33}$$

we can use Gershgorin's Theorem [3] to obtain the bound

$$0 \le \left(\frac{|\omega|\Delta x}{c}\right)^2 \le 4, \tag{4.34}$$

i.e., $\max|\omega| = \rho(A) \le 2c/\Delta x$. Assuming equality here,[7] the convergence factor is $\rho(\mathbf{g}'(\mathbf{v})) = 2c\gamma\Delta t/\Delta x$ [c.f. discussion following (3.16)] and the iteration converges if and only if

$$\frac{c\Delta t}{\Delta x} < \frac{1}{2\gamma}. \tag{4.35}$$

This has the same form as a CFL condition for explicit time differencing, showing that the simple fixed-point iteration defeats the purpose of semi-implicit time differencing.

The iteration (4.29) can be viewed as a "simultaneous" iteration scheme, since it updates $\mathbf{u}$ and $\mathbf{h}$ at the same time (like Jacobi iteration). Alternatively, we could use the "successive" iteration

$$\begin{aligned} \mathbf{u}^{(k+1)} &= \mathbf{f}_u + \frac{g\gamma\Delta t}{\Delta x} D\mathbf{h}^{(k)}, \\ \mathbf{h}^{(k+1)} &= \mathbf{f}_h - \frac{\bar{h}\gamma\Delta t}{\Delta x} D^T\mathbf{u}^{(k+1)}, \end{aligned} \tag{4.36}$$

which updates $\mathbf{u}$ first and then $\mathbf{h}$ (like Gauss-Seidel). This can be written in matrix form as

$$\mathbf{v}^{(k+1)} = \widetilde{\mathbf{g}}(\mathbf{v}^{(k)}) := \widetilde{\mathbf{f}} + \widetilde{A}\mathbf{v}^{(k)} \tag{4.37}$$

where

$$\widetilde{\mathbf{f}} = \begin{bmatrix} \mathbf{f}_u \\ \mathbf{f}_h - \dfrac{\bar{h}\gamma\Delta t}{\Delta x} D^T\mathbf{f}_u \end{bmatrix}, \qquad \widetilde{A} = \begin{bmatrix} 0 & \dfrac{g\gamma\Delta t}{\Delta x} D \\ 0 & -\left(\dfrac{c\gamma\Delta t}{\Delta x}\right)^2 D^T D \end{bmatrix}. \tag{4.38}$$

Proceeding as above, we find the convergence factor is $\rho(\widetilde{\mathbf{g}}'(\mathbf{v})) = \rho(\widetilde{A}) = (2c\gamma\Delta t/\Delta x)^2 = (\rho(\mathbf{g}'(\mathbf{v})))^2$, so the successive iteration converges twice as fast as the simultaneous iteration—when it converges. However, the convergence criterion is still (4.35), thus again defeating the purpose of semi-implicit time differencing.

[7] For a more precise analysis we can use the actual eigenvalues of (4.33) in place of the bound (4.34) to show that $\rho(A) \to 2c/\Delta x$ as $N \to \infty$.

4.2.3 Other iterative schemes

To investigate other iterative schemes it is convenient to expand the implicit problem (4.27) as

$$\begin{aligned} \mathbf{u} - \frac{g\gamma\Delta t}{\Delta x} D\mathbf{h} &= \mathbf{f}_u, \\ \mathbf{h} + \frac{\bar{h}\gamma\Delta t}{\Delta x} D^T\mathbf{u} &= \mathbf{f}_h \end{aligned} \tag{4.39}$$

and eliminate $\mathbf{u}$ to obtain

$$\left(I + C^2 D^T D\right)\mathbf{h} = \mathbf{f}_h - \frac{\bar{h}\gamma\Delta t}{\Delta x} D^T \mathbf{f}_u, \tag{4.40}$$

where $C = c\gamma\Delta t/\Delta x$ is the Courant number (scaled by γ) and $I \in \mathbb{R}^{N\times N}$ is the identity matrix. The matrix on the left side of (4.40) is

$$\begin{bmatrix} 1+C^2 & -C^2 & & & \\ -C^2 & 1+2C^2 & -C^2 & & \\ & \ddots & \ddots & \ddots & \\ & & -C^2 & 1+2C^2 & -C^2 \\ & & & -C^2 & 1+C^2 \end{bmatrix} \tag{4.41}$$

which is strictly diagonally dominant (for any $\Delta t > 0$), reflecting the fact that (4.40) is a discretized form of the elliptic problem for h. Thus, all of the classical iterative methods (e.g., Jacobi, Gauss-Seidel) converge for any Δt. For example, the Jacobi iteration matrix for (4.40) is

$$M_J = \begin{bmatrix} 0 & \alpha & & & \\ \beta & 0 & \beta & & \\ & \ddots & \ddots & \ddots & \\ & & \beta & 0 & \beta \\ & & & \alpha & 0 \end{bmatrix} \tag{4.42}$$

where $\alpha = C^2/(1+C^2)$ and $\beta = C^2/(1+2C^2)$. By Gershgorin's Theorem we have $\rho(M_J) \leq \max(\alpha, 2\beta) = 2\beta < 1$ so the Jacobi iteration converges.

4.3 Two-dimensional shallow-water equations

Turning now to the nonlinear shallow water equations in two dimensions, we recall that we need to treat the height gradient term (and possibly the Coriolis term) in the momentum equation and the divergence term in the continuity equation implicitly. Here we illustrate how this can be done in both the momentum form (4.1)–(4.2) and the vorticity-divergence form.

4.3.1 Momentum form

To split the terms in the shallow water model, we specify a positive reference height $\bar{h}$ and set $h = \bar{h} + h'$. Likewise, we set $f = \bar{f} + f'$ so we can specify $\bar{f} = 0$ or $\bar{f} = f$ to treat the Coriolis term explicitly or implicitly, respectively. Note that $\bar{h}$ and $\bar{f}$ must be constant in time but not necessarily in space. Then the shallow-water equations become

$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial t} + \bar{f}\mathbf{k} \times \mathbf{v} + g\nabla h &= \mathbf{F} - \mathbf{v} \cdot \nabla \mathbf{v} - f'\mathbf{k} \times \mathbf{v}, && (4.43) \\ \frac{\partial h}{\partial t} + \nabla \cdot (\bar{h}\mathbf{v}) &= Q - \nabla \cdot (h'\mathbf{v}). && (4.44) \end{aligned}$$

Treating the terms on the left implicitly and those on the right explicitly leads to an implicit system of the form

$$\begin{aligned} \mathbf{v} + \tau\bar{f}\mathbf{k} \times \mathbf{v} + \tau g\nabla h &= \mathbf{V}, && (4.45) \\ h + \tau\nabla \cdot (\bar{h}\mathbf{v}) &= H, && (4.46) \end{aligned}$$

where $\mathbf{v}$ and h are the variables at the new time level, $\mathbf{V}$ and H are computed explicitly from values at the old time level, and (for simplicity of notation) $\tau = \gamma\Delta t$, where γ is a constant determined by the implicit scheme chosen [c.f. (3.1)]. Taking the cross product of $\tau\bar{f}\mathbf{k}$ with (4.45) and subtracting this from (4.45) allows us to solve for $\mathbf{v}$ in terms of h, resulting in

$$\mathbf{v} = a^2 \left[\left(\mathbf{V} - \tau\bar{f}\mathbf{k} \times \mathbf{V} \right) - \tau g \left(\nabla h - \tau\bar{f}\mathbf{k} \times \nabla h \right) \right] \qquad (4.47)$$

where $a = \left(1 + \tau^2\bar{f}^2\right)^{-1/2}$ (usually $|\tau\bar{f}| \ll 1$ so $a \approx 1$). Substituting this into (4.46) yields

$$h + \tau^2 \mathcal{J}\left(h, \bar{c}^2\tau\bar{f}\right) - \tau^2\nabla \cdot \left(\bar{c}^2\nabla h\right) = G, \qquad (4.48)$$

where $\bar{c} = a\left(g\bar{h}\right)^{1/2}$, $\mathcal{J}(\cdot,\cdot)$ is the Jacobian operator, and

$$G = H - \tau\nabla \cdot \left[a^2\bar{h}\left(\mathbf{V} - \tau\bar{f}\mathbf{k} \times \mathbf{V}\right)\right]. \qquad (4.49)$$

Since $\bar{c}$ and $\bar{f}$ are known and $\bar{c} \neq 0$, (4.48) is a linear elliptic equation for h. It may also be written in the form

$$h - \tau^2\mathbf{b} \cdot \nabla h - \tau^2\bar{c}^2\nabla^2 h = G \qquad (4.50)$$

where

$$\mathbf{b} = \nabla\left(\bar{c}^2\right) + \mathbf{k} \times \nabla\left(\tau\bar{f}\bar{c}^2\right). \qquad (4.51)$$

If $\bar{h}$ does not vary too rapidly in space, the first-order term in (4.48) [or (4.50)] will be relatively small so the second-order term will dominate; thus we should be able to solve this equation for h using a standard iterative (or multigrid) method. The corresponding velocity $\mathbf{v}$ then can be computed from (4.47).

Several special cases are of interest. First, if $\bar{f} = 0$ (i.e., the Coriolis term is treated explicitly) then $a = 1$, $\bar{c} = \left(g\bar{h}\right)^{1/2}$, (4.47) reduces to

$$\mathbf{v} = \mathbf{V} - \tau g\nabla h, \qquad (4.52)$$

and (4.48) reduces to

$$h - \tau^2 \nabla \cdot \left(\bar{c}^2 \nabla h \right) = G, \tag{4.53}$$

where

$$G = H - \tau \nabla \cdot \left(\bar{h} \mathbf{V} \right). \tag{4.54}$$

If, in addition, $\bar{h}$ is constant in space, then (4.53) further reduces to the (modified) Helmholtz equation

$$h - \bar{c}^2 \tau^2 \nabla^2 h = G. \tag{4.55}$$

Solving (4.55) by an iterative or multigrid method is straightforward, and (4.53) is only a slight generalization. Second, if $\bar{f}$ is constant (i.e., an f-plane) and $\bar{h}$ is constant, then again we obtain the simple (modified) Helmholtz equation (4.55), even if the Coriolis term is treated implicitly.

4.3.2 Vorticity-divergence form

The momentum form momentum form (4.1)–(4.2) treated above has the disadvantage that the components of the velocity $\mathbf{v}$ are not true scalars (their values depend on the coordinate system chosen). Thus, Heikes and Randall [9] and others advocate using vorticity and divergence instead. To do so, we dot $\mathbf{k}$ with the curl of the momentum equation (4.43) to obtain the vorticity equation

$$\frac{\partial \eta}{\partial t} + \nabla \cdot (\eta \mathbf{v}) = \mathbf{k} \cdot \nabla \times \mathbf{F}, \tag{4.56}$$

where $\eta = f + \zeta$ is the absolute vorticity and $\zeta = \mathbf{k} \cdot \nabla \times \mathbf{v}$ is the relative vorticity. Likewise, taking the divergence of the momentum equation gives the divergence equation

$$\frac{\partial \delta}{\partial t} - \mathbf{k} \cdot \nabla \times (\eta \mathbf{v}) + \nabla^2 (gh + K) = \nabla \cdot \mathbf{F}, \tag{4.57}$$

where $\delta = \nabla \cdot \mathbf{v}$ is the divergence and $K = \frac{1}{2} \mathbf{v} \cdot \mathbf{v}$ is the (specific) kinetic energy. Since we must be able to compute the velocity from the vorticity and divergence, we introduce the streamfunction ψ and velocity potential χ satisfying

$$\mathbf{v} = \nabla \chi + \mathbf{k} \times \nabla \psi \tag{4.58}$$

so that

$$\nabla^2 \psi = \zeta, \qquad \nabla^2 \chi = \delta. \tag{4.59}$$

It is convenient to use (4.58) to eliminate $\mathbf{v}$ from (4.56), (4.57), and (4.2), yielding

$$\begin{aligned}
\frac{\partial \eta}{\partial t} + \nabla \cdot (\eta \nabla \chi) - \mathcal{J}(\eta, \psi) &= \mathbf{k} \cdot \nabla \times \mathbf{F}, && (4.60) \\
\frac{\partial \delta}{\partial t} - \nabla \cdot (\eta \nabla \psi) - \mathcal{J}(\eta, \chi) + \nabla^2 (gh + K) &= \nabla \cdot \mathbf{F}, && (4.61) \\
\frac{\partial h}{\partial t} + \nabla \cdot (h \nabla \chi) - \mathcal{J}(h, \psi) &= Q. && (4.62)
\end{aligned}$$

In this form the only operators needed are the flux divergence, Laplacian, and Jacobian, since we can also write the kinetic energy as

$$K = \frac{1}{2}\left[\nabla\cdot(\chi\nabla\chi) - \chi\nabla^2\chi + \nabla\cdot(\psi\nabla\psi) - \psi\nabla^2\psi\right] + \mathcal{J}(\psi,\chi). \tag{4.63}$$

To split the terms for semi-implicit time differencing, we again write $h = \bar{h} + h'$ and $f = \bar{f} + f'$ (where $\bar{h}$ and $\bar{f}$ are specified and constant in time but not necessarily in space), and set $\eta = \bar{f} + \eta'$ so $\eta' = f' + \zeta$. In particular, we can set $\bar{f} = 0$ or $\bar{f} = f$ to treat the Coriolis terms explicitly or implicitly, respectively. Then we can write (4.60)–(4.62) in the form

$$\begin{aligned}
\frac{\partial\eta}{\partial t} + \nabla\cdot\left(\bar{f}\nabla\chi\right) - \mathcal{J}\left(\bar{f},\psi\right) &= \mathbf{k}\cdot\nabla\times\mathbf{F} - \nabla\cdot(\eta'\nabla\chi) + \mathcal{J}(\eta',\psi), && (4.64)\\
\frac{\partial\delta}{\partial t} - \nabla\cdot\left(\bar{f}\nabla\psi\right) - \mathcal{J}\left(\bar{f},\chi\right) + g\nabla^2 h &= \nabla\cdot\mathbf{F} + \nabla\cdot(\eta'\nabla\psi) + \mathcal{J}(\eta',\chi) - \nabla^2 K, && (4.65)\\
\frac{\partial h}{\partial t} + \nabla\cdot\left(\bar{h}\nabla\chi\right) - \mathcal{J}\left(\bar{h},\psi\right) &= Q - \nabla\cdot(h'\nabla\chi) + \mathcal{J}(h',\psi), && (4.66)
\end{aligned}$$

Treating the terms on the left implicitly and the the terms on the right explicitly leads to an implicit problem of the form

$$\begin{aligned}
\zeta + \tau\nabla\cdot(\bar{f}\nabla\chi) - \tau\mathcal{J}\left(\bar{f},\psi\right) &= Z, && (4.67)\\
\delta - \tau\nabla\cdot(\bar{f}\nabla\psi) - \tau\mathcal{J}\left(\bar{f},\chi\right) + \tau g\nabla^2 h &= D, && (4.68)\\
h + \tau\nabla\cdot(\bar{h}\nabla\chi) - \tau\mathcal{J}\left(\bar{h},\psi\right) &= H, && (4.69)
\end{aligned}$$

where ψ, χ, and h are values at the new time level, ζ and δ satisfy (4.59), and Z, D, and H are computed explicitly from values at the old time level. The method of solving the implicit system (4.67)–(4.69) and (4.59) depends on what we assume for $\bar{h}$ and $\bar{f}$.

First, if we take $\bar{f} = 0$ (i.e., treat the Coriolis terms explicitly), then (4.67) and (4.68) reduce to

$$\zeta = Z \tag{4.70}$$

and

$$\delta + \tau g\nabla^2 h = D, \tag{4.71}$$

to be solved together with (4.69). Setting

$$\hat{\chi} = \chi + \tau g h \tag{4.72}$$

and using (4.59) we can write (4.71) as

$$\nabla^2\hat{\chi} = D. \tag{4.73}$$

Then solving the Poisson problems (4.59) and (4.73) for ψ and $\hat{\chi}$, we can substitute in (4.69) from (4.72) to obtain

$$h - \tau^2\nabla\cdot\left(\bar{c}^2\nabla h\right) = G \tag{4.74}$$

[cf. (4.53)], where

$$G = H - \tau\nabla\cdot\left(\bar{h}\nabla\hat{\chi}\right) + \tau\mathcal{J}\left(\bar{h},\psi\right). \tag{4.75}$$

Once this linear elliptic equation is solved for h, we can compute χ from (4.72). Note that if $\bar{h}$ is constant in space, (4.74) reduces to the (modified) Helmholtz equation (4.55) as in the momentum form, and we could avoid introducing $\hat{\chi}$ if desired.

Second, if we take $\bar{f}$ to be nonzero but constant (i.e., an f-plane with the Coriolis terms treated implicitly) and take $\bar{h}$ to be constant, then (4.67)–(4.69) reduce to

$$\begin{aligned} \zeta + \tau\bar{f}\delta &= Z, && (4.76)\\ \delta - \tau\bar{f}\zeta + \tau g\nabla^2 h &= D, && (4.77)\\ h + \tau\bar{h}\delta &= H. && (4.78) \end{aligned}$$

Eliminating ζ between (4.76) and (4.77) yields

$$\delta + a^2\tau g\nabla^2 h = a^2\left(D + \tau\bar{f}Z\right) \tag{4.79}$$

where $a = \left(1+\tau^2\bar{f}^2\right)^{-1/2}$ as before. Then eliminating δ between (4.78) and (4.79) leads to the (modified) Helmholtz problem

$$h - \bar{c}^2\tau^2\nabla^2 h = G \tag{4.80}$$

[cf. (4.55)], where $\bar{c} = a\left(g\bar{h}\right)^{1/2}$ as before and

$$G = H - a^2\tau\bar{h}\left(D + \tau\bar{f}Z\right). \tag{4.81}$$

Once (4.80) is solved for h we can compute δ from (4.79) and ζ from (4.76), and then compute ψ and χ by solving the Poisson problems (4.59).

Finally, with no simplifying assumptions on $\bar{h}$ and $\bar{f}$, we can solve the implicit system (4.67)–(4.69) and (4.59) by essentially converting it to the momentum form as follows. The key is to introduce the "forced velocity"

$$\mathbf{V} = \nabla X + \mathbf{k}\times\nabla\Psi, \tag{4.82}$$

which we can compute from Z and D by solving the Poisson problems

$$\nabla^2\Psi = \mathbf{k}\cdot\nabla\times\mathbf{V} = Z, \qquad \nabla^2 X = \nabla\cdot\mathbf{V} = D. \tag{4.83}$$

Then the velocity $\mathbf{v}$ derived from ψ, χ, and h satisfying the implicit system (4.67)–(4.69) satisfies (4.45) and (4.46), so we can solve (4.48) for h and compute $\mathbf{v}$ from (4.47), as explained for the momentum form above. From $\mathbf{v}$ we can then obtain ψ and χ using (4.58) and solving the Poisson problems (4.59).

In summary, treating the shallow-water equations with a semi-implicit method leads to a single linear elliptic problem to be solved for h at the new time level; this problem should be solvable by standard iterative or multigrid methods. In the momentum form the corresponding velocity $\mathbf{v}$ is computed directly from h. The vorticity-divergence form requires solving two more Poisson problems to obtain the corresponding streamfunction ψ and velocity potential χ (plus the two additional Poisson problems (4.83) if the Coriolis terms are treated implicitly and either $\bar{f}$ or $\bar{h}$ is not constant in space).

5 Convergence Analysis

Turning now to the analysis of semi-implicit schemes, we first consider the question of *convergence*: if we fix t and use n time steps to compute $\psi^n \approx \psi(t)$ with $\Delta t = t/n$, does ψ^n converge to $\psi(t)$ as $n \to \infty$, i.e., as $\Delta t \to 0$? Section 2 reviewed the principal convergence results available in the literature; in this section we fill in some of the details. Here we follow the analysis of [2] and [11].

5.1 Combined linear multistep methods

In what follows, we will concentrate on semi-implicit schemes which treat both the implicit and explicit terms using linear multistep methods. The most general such *combined linear multistep (CLM) method* for the problem (2.1) can be written in the form

$$\frac{1}{\Delta t}\sum_{j=0}^{m} c_j \psi^{n+1-j} = \sum_{j=0}^{m} a_j \mathcal{A}\left(\psi^{n+1-j}\right) + \sum_{j=0}^{m} b_j \mathcal{B}\left(\psi^{n+1-j}\right). \tag{5.1}$$

The scheme is completely specified by the number m of steps and the constants $\mathbf{a} = (a_0, a_1, \ldots, a_m)$, $\mathbf{b} = (b_0, b_1, \ldots, b_m)$, and $\mathbf{c} = (c_0, c_1, \ldots, c_m)$. Note that there is an extra degree of freedom that may be removed by specifying a normalization condition on c_0, which must be nonzero; e.g., $c_0 = 1$. We also assume that at least one of a_m, b_m, c_m is nonzero (so the scheme actually uses m steps). The scheme (5.1) is semi-implicit if $a_0 \neq 0$ and $b_0 = 0$, which we will assume; the resulting method is a combination of an implicit m-step method given by $\mathbf{c}$ and $\mathbf{a}$ and an explicit m-step method given by $\mathbf{c}$ and $\mathbf{b}$. Examples include the backward/forward scheme (2.2), for which

$$m = 1, \qquad \mathbf{c} = (1, -1), \qquad \mathbf{a} = (1, 0), \qquad \mathbf{b} = (0, 1), \tag{5.2}$$

the trapezoidal/leapfrog scheme (2.3), for which

$$m = 2, \qquad \mathbf{c} = \left(\tfrac{1}{2}, 0, -\tfrac{1}{2}\right), \qquad \mathbf{a} = \left(\tfrac{1}{2}, 0, \tfrac{1}{2}\right), \qquad \mathbf{b} = (0, 1, 0), \tag{5.3}$$

and the semi-implicit AB2 scheme (2.4), for which

$$m = 2, \qquad \mathbf{c} = (1, -1, 0), \qquad \mathbf{a} = (\theta, 1-\theta, 0), \qquad \mathbf{b} = \left(0, \tfrac{3}{2}, -\tfrac{1}{2}\right). \tag{5.4}$$

Each time step of the scheme (5.1) requires solving the implicit problem

$$c_0\psi^{n+1} - a_0\Delta t\mathcal{A}\left(\psi^{n+1}\right) = \sum_{j=1}^{m}\left[-c_j\psi^{n+1-j} + a_j\Delta t\mathcal{A}\left(\psi^{n+1-j}\right) + b_j\Delta t\mathcal{B}\left(\psi^{n+1-j}\right)\right], \tag{5.5}$$

which has the form of (3.1) with $\gamma = a_0/c_0$. Thus, the scheme requires solving one implicit problem per time step. In general, the storage required is one location for each component of $\mathcal{A}(\psi)$, $\mathcal{B}(\psi)$, and ψ for each nonzero value of a_j, b_j, and c_j, respectively, for $j = 1, 2, \ldots, m$ (and possibly more if intervening coefficients are zero). However, in some cases certain combinations of these values might be stored instead (or $\mathcal{A}$ or $\mathcal{B}$ recomputed from ψ) to reduce the storage.

5.2 Consistency, order, and truncation error

To be of any value, the scheme (5.1) must approximate the equation (2.1). This is measured by the *truncation error* τ, defined as the amount by which the true solution ψ of (2.1) fails to satisfy the discrete equation (5.1). Assuming that $\psi \in C^{p+1}$ and $\mathcal{A}, \mathcal{B} \in C^p$ we can use Taylor expansions to show

$$\begin{aligned} \tau(t;\Delta t) &:= \frac{1}{\Delta t}\sum_{j=0}^{m} c_j \psi(t-j\Delta t) - \sum_{j=0}^{m} a_j \mathcal{A}\left(\psi(t-j\Delta t)\right) - \sum_{j=0}^{m} b_j \mathcal{B}\left(\psi(t-j\Delta t)\right) \\ &= \frac{1}{\Delta t}\left(\sum_{j=0}^{m} c_j\right)\psi(t) + \sum_{k=0}^{p}\left[\frac{1}{(k+1)!}\left(\sum_{j=0}^{m}(-j)^{k+1}c_j\right)\psi^{(k+1)}(t)\right. \\ &- \left.\frac{1}{k!}\left(\sum_{j=0}^{m}(-j)^k a_j\right)\mathcal{A}^{(k)}(t) - \frac{1}{k!}\left(\sum_{j=0}^{m}(-j)^k b_j\right)\mathcal{B}^{(k)}(t)\right](\Delta t)^k + o\left((\Delta t)^p\right) \end{aligned} \tag{5.6}$$

where $\mathcal{A}^{(k)}(t) = (d^k/dt^k)A(\psi(t))$ and similarly for $\mathcal{B}^{(k)}(t)$. The method is *consistent* if $\tau \to 0$ as $\Delta t \to 0$ with t fixed; from (5.6) this requires at least

$$\sum_{j=0}^{m} c_j = 0. \tag{5.7}$$

The method is of *order* (at least) p if $\tau = O\left((\Delta t)^p\right)$ for some integer $p > 0$; using the kth derivative of (2.1) in (5.6) shows that this requires

$$\frac{1}{(k+1)!}\sum_{j=0}^{m}(-j)^{k+1}c_j = \frac{1}{k!}\sum_{j=0}^{m}(-j)^k a_j = \frac{1}{k!}\sum_{j=0}^{m}(-j)^k b_j \tag{5.8}$$

for $k = 0, 1, \ldots, p-1$. Since consistency is equivalent to order at least $p = 1$, for the method to be consistent the conditions (5.8) with $k = 0$, i.e.,

$$-\sum_{j=0}^{m} jc_j = \sum_{j=0}^{m} a_j = \sum_{j=0}^{m} b_j \tag{5.9}$$

must hold in addition to (5.7). If the method is of order p then the truncation error reduces to

$$\begin{aligned} \tau(t;\Delta t) &= \sum_{j=0}^{m}\left[\frac{(-j)^{p+1}}{(p+1)!}c_j - \frac{(-j)^p}{p!}a_j\right]\mathcal{A}^{(p)}(t)(\Delta t)^p \\ &+ \sum_{j=0}^{m}\left[\frac{(-j)^{p+1}}{(p+1)!}c_j - \frac{(-j)^p}{p!}b_j\right]\mathcal{B}^{(p)}(t)(\Delta t)^p + o\left((\Delta t)^p\right). \end{aligned} \tag{5.10}$$

From (5.10) we draw the following conclusions [cf. section 2]:

- If the implicit and explicit methods are consistent separately, then the resulting semi-implicit method (5.1) is consistent.
- The order of the semi-implicit method is the minimum of the orders of the implicit and explicit methods considered separately.
- If the implicit method has the lower order, then the leading term(s) in the truncation error are proportional to derivatives of $\mathcal{A}(\psi)$ and do not involve $\mathcal{B}(\psi)$.

5.3 Zero-stability and convergence

Another property the scheme must have to be useful is *zero-stability.* To understand this, we note that if the problem (2.1) is well-posed, then for any fixed time t the solution $\psi(t)$ depends continuously on the initial data $\psi(0)$. The scheme (5.1) is *zero-stable* if it has this same property *independent of the step size* Δt, i.e., if the sensitivity of the discrete solution ψ^n at fixed $t = n\Delta t$ to changes in the data $\psi(0)$ is bounded independent of Δt as $\Delta t \to 0$ ($n \to \infty$). It can be shown (e.g., [18]) that the method is zero-stable if and only if the roots $\{r_j\}_{j=1}^m$ of the *characteristic polynomial*

$$P_0(z) := \sum_{j=0}^{m} c_j z^{m-j} \tag{5.11}$$

satisfy the *root condition*

$$|r_j| \leq 1 \quad \text{and} \quad |r_j| = 1 \text{ only if } r_j \text{ is a simple root.} \tag{5.12}$$

Note that zero-stability depends only on the constants c_j. Also, for a consistent method $P_0(1) = 0$ by (5.7), so we can always index the roots so that $r_1 = 1$.

The notion of zero-stability employed here differs from the notion of *absolute stability* to be treated in the next section. The latter is associated with a fixed step size $\Delta t > 0$, while the former deals with the limit as $\Delta t \to 0$. Zero-stability is sometimes referred to as *Lax-stability* due to its role in the Lax-Richtmyer equivalence theorem, which states that the method (5.1) is *convergent* if and only if it is both *consistent* and *stable.* Normally stated and proved for conventional methods for initial value problems (e.g., [18]), this theorem also holds for the case of CLM methods [11]. Thus, zero-stability is essential: a method unstable in this sense will not in general produce results which converge (as $\Delta t \to 0$) to the proper solution. Furthermore, as we will see in the next section, zero-stability is a limiting case of the more stringent requirement of absolute stability.

6 Absolute Stability Analysis

In the previous section we addressed the question of whether or not a semi-implicit scheme converges, i.e., gives the right answer at a *fixed time* t in the limit as $\Delta t = t/n \to 0$ (or equivalently as $n \to \infty$). Certainly one should not use a scheme without this property. However, in practice one uses a fixed $\Delta t \neq 0$ (in fact, the largest Δt which gives sufficiently accurate results), so a subsequent question essential to practical applications remains: is the scheme *absolutely stable*? By this we mean: if the *time step* Δt *is fixed*, does the solution remain bounded as $t = n\Delta t \to \infty$ (or equivalently as $n \to \infty$)? Indeed, the whole point of semi-implicit time differencing is to increase the allowable time step for efficiency. While convergence—and the related notion of zero-stability—can be analyzed independent of the problem (subject only to some reasonable conditions on smoothness of the solution), absolute stability depends not only on the method but also on the problem to which it is applied and on the time step. For example, the leapfrog scheme is absolutely stable for oscillatory problems (conditionally so: stable only with sufficiently small time step) but absolutely unstable for dissipative problems (unconditionally so: unstable with any time step).

While the question of convergence of semi-implicit schemes has been answered in considerable generality (see section 5), there is less known about their absolute stability. Ascher et al. [2] analyzed the stability of selected semi-implicit LM schemes for convection-diffusion problems (where the diffusion is treated implicitly), and Kang [11] likewise treated semi-implicit RK methods up to order 3. However, for meteorological applications one is often interested in treating oscillatory components (fast waves) implicitly, and there seems to be less known about stability in this context. Kang [11] found no semi-implicit RK method (to third order) to be absolutely stable for oscillatory problems, and numerical experiments suggest the same negative result for the semi-implicit fourth-order RK method used in [8]. On the other hand, some semi-implicit LM methods (e.g., trapezoidal/leapfrog) have been used successfully in many atmospheric models, and certain schemes have been analyzed in the context of specific models. In this section we give the machinery needed to study absolute stability of semi-implicit schemes and use it to search for practical methods.

6.1 Test equation

To analyze the absolute stability of conventional (not semi-implicit) schemes one uses a simple scalar test equation. To motivate this, suppose the equations have been discretized in space so $\psi(t) = \mathbf{v}(t) \in \mathbb{R}^N$ for some N. In the linear constant-coefficient case where $\mathcal{A}(\psi(t)) = A\mathbf{v}$ for a constant matrix $A \in \mathbb{R}^{N\times N}$, the equation to be solved is

$$\frac{d\mathbf{v}}{dt} = A\mathbf{v}. \tag{6.1}$$

If A is diagonalizable, i.e., there exists a nonsingular matrix P such that $\Lambda = P^{-1}AP$ is diagonal, then the transformation $\mathbf{w} = P^{-1}\mathbf{v}$ transforms (6.1) into

$$\frac{d\mathbf{w}}{dt} = \Lambda\mathbf{w}, \tag{6.2}$$

in which the components are decoupled. The resulting solution $\mathbf{w}(t) = e^{\Lambda t}\mathbf{w}(0)$ is then bounded as $t \to \infty$ provided the eigenvalues $\lambda_1, \ldots, \lambda_N$ (which are the diagonal entries of Λ) all have non-positive real parts. Thus, one analyzes the absolute stability of a conventional scheme by applying it to the scalar test equation

$$\frac{d\psi}{dt} = \lambda\psi \qquad (\lambda \in \mathbb{C}). \tag{6.3}$$

This represents a single component of (6.2), with λ an eigenvalue of the matrix A. Note that imaginary and negative real values of λ correspond to oscillatory and decaying solutions, respectively, and thus to wave and dissipative processes.

The connection established above between the linear problem (6.1) and the test equation (6.3) is precise only when with A constant and diagonalizable. Even in that case, absolute stability may not be enough. To see this, note that the solution of (6.1) is $\mathbf{v}(t) = Pe^{\Lambda t}P^{-1}\mathbf{v}(0)$, so

$$\|\mathbf{v}(t)\| \leq \|P\| \left\|e^{\Lambda t}\right\| \left\|P^{-1}\right\| \|\mathbf{v}(0)\| \tag{6.4}$$

in any vector norm and corresponding operator matrix norm. Thus, for a stable problem (all eigenvalues with non-positive real parts),

$$\|\mathbf{v}(t)\| \leq \operatorname{cond}(P) \|\mathbf{v}(0)\| \tag{6.5}$$

where $\operatorname{cond}(P) := \|P\| \|P^{-1}\|$ is the *condition number* of P (which is always at least 1). Now P can be chosen to be orthogonal ($P^*P = I$) if and only if A is *normal* ($A^*A = AA^*$); in this case $\operatorname{cond}(P) = 1$ in the l_2 norm, so the solution is not only bounded but satisfies $\|\mathbf{v}(t)\| \leq \|\mathbf{v}(0)\|$, which guarantees that any perturbations introduced will remain small. However, if A is not normal then the effect of any perturbations may be large (magnified by $\operatorname{cond}(P)$, which may be large). One might hope that if the problem is in some sense "close to normal" then absolute stability for the test equation (6.3) would at least suggest that the method will work acceptably; however, even in the linear case this need not be true if A is not diagonalizable or not constant (e.g., [1], page 34), and in nonlinear problems all bets are off (e.g., [10]). Nevertheless, this linear analysis in terms of the eigenvalues of the matrix A is in general the best we can do.

For a semi-implicit scheme the analysis is more complicated, due to the need to represent two terms in the equations: we need a generalization of the test equation (6.3). Thus, consider the linear constant-coefficient case where $\mathcal{A}(\psi(t)) = A\mathbf{v}$ and $\mathcal{B}(\psi(t)) = B\mathbf{v}$ for constant matrices $A, B \in \mathbb{R}^{N \times N}$. The equation (2.1) to be solved is then

$$\frac{d\mathbf{v}}{dt} = A\mathbf{v} + B\mathbf{v}. \tag{6.6}$$

If A and B are *simultaneously* diagonalizable, i.e., if there exists a nonsingular matrix P such that $\Lambda_A = P^{-1}AP$ and $\Lambda_B = P^{-1}BP$ are both diagonal, then the transformation $\mathbf{w} = P^{-1}\mathbf{v}$ transforms (6.6) into

$$\frac{d\mathbf{w}}{dt} = \Lambda_A\mathbf{w} + \Lambda_B\mathbf{w}, \tag{6.7}$$

in which the components are decoupled. Thus, one analyzes the absolute stability of a semi-implicit scheme by applying it to the generalized scalar test equation

$$\frac{d\psi}{dt} = a\psi + b\psi \qquad (a, b \in \mathbb{C}). \tag{6.8}$$

This represents a single component of (6.2), with a and b representing eigenvalues of A and B, respectively. As above, imaginary and negative real values of a or b correspond to oscillatory and decaying solutions, respectively, and thus to wave and dissipative processes.

It can be shown (e.g., [14]) that A and B are simultaneously diagonalizable if and only if they are diagonalizable and commute (i.e., $AB = BA$). While this requirement is stringent, it is satisfied, for example, by the linearized shallow-water equations studied in section 4.1. There we found that if we treat the gravity-wave and Coriolis terms implicitly and the advective terms explicitly, the resulting matrices are simultaneously diagonalizable (indeed, even orthogonally so). The eigenvalues of the implicit operator A are 0 and $\pm i\nu$ (where ν is the gravity-inertia wave frequency); the only eigenvalue of the explicit operator B is $i\bar{\omega}$ (the frequency corresponding to advection). Thus, we will be especially interested in the case where $a = i\omega_f$ and $b = i\omega_s$, with ω_f and ω_s representing the (real) frequencies of fast and slow waves in a model.

6.2 Absolute stability regions

Any time differencing scheme applied to a particular problem for $\psi(t)$ such as (2.1) generates a solution $\{\psi^n\}_{n=0}^{\infty}$; if for a fixed time step Δt, ψ^n is bounded as $n \to \infty$, we say that the scheme is *absolutely stable* for that problem and time step. In the case of a conventional scheme applied to the scalar test equation (6.3), the problem is completely characterized by the complex number λ, so we can ask whether the scheme is absolutely stable for a given λ and Δt. It turns out that the solution ψ^n depends only on the product $\lambda\Delta t$, so we can define the *region of absolute stability* as the set of values of $z = \lambda\Delta t$ in the complex z-plane for which the scheme is absolutely stable when applied to (6.3) with the time step Δt. Recall that pure imaginary and negative real values of λ correspond to oscillatory and dissipative processes, respectively. It is reasonable to hope for absolute stability only for $\mathrm{Re}\,(\lambda\Delta t) \leq 0$, for which the original problem has bounded solutions. The regions of absolute stability[8] for the Forward (FOR), second-order Adams-Bashforth (AB2), Backward (BACK), and Trapezoidal (TRAP) schemes are shown in Fig. 3. Note that the FOR and AB2 schemes (which are explicit) are unconditionally *unstable* for oscillatory processes (imaginary axis). In contrast, the BACK and TRAP schemes (which are implicit) are stable not just for oscillatory processes but also for the whole left half-plane $\mathrm{Re}\,(\lambda\Delta t) \leq 0$; methods with this property are said to be *A-stable*.

To analyze a semi-implicit method we must use (6.8) as explained above. For this problem it turns out that the absolute stability depends on the products $\alpha = a\Delta t$ and $\beta = b\Delta t$.

[8] Computed as described below for semi-implicit methods with $\beta = 0$.

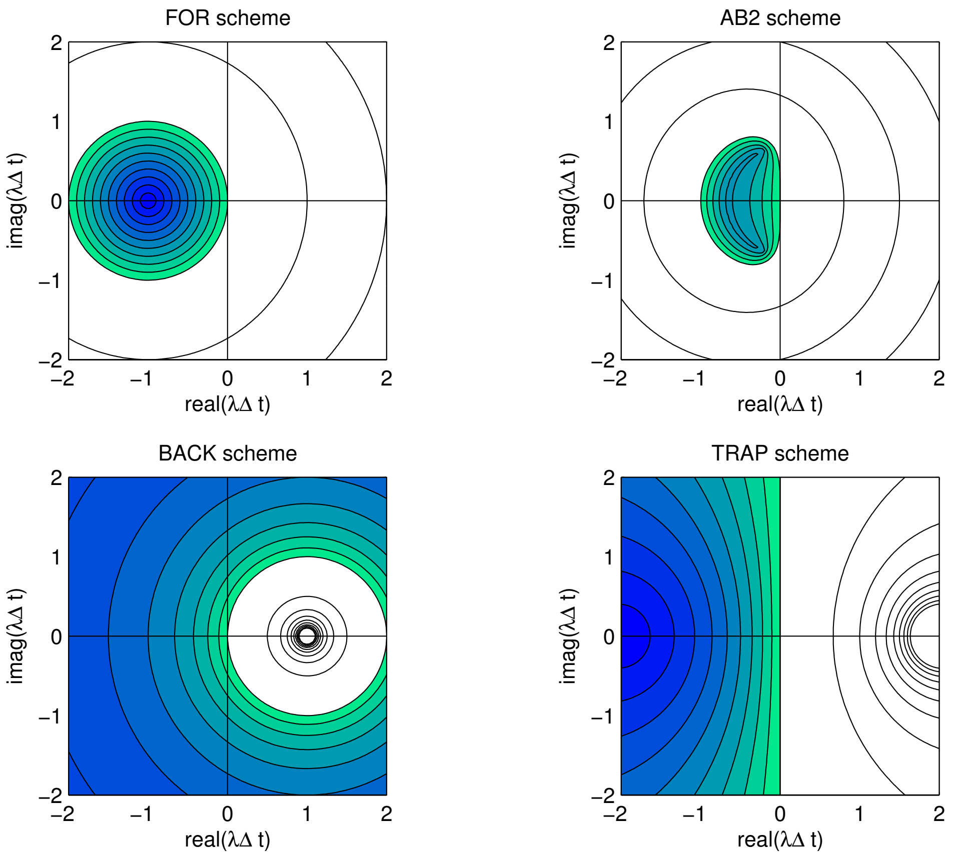

Figure 2: Absolute stability regions for four one-step schemes. The stability regions are shaded, and the contour interval for the amplification factors $r = \max_j |r_j|$ are 0.1 for $|r| \leq 1$ and and 1.0 for $|r| > 1$.

Specifically, any one-step semi-implicit method applied to (6.8) yields a solution of the form

$$\psi^n = [r(\alpha, \beta)]^n \psi^0 \tag{6.9}$$

where the complex number r is the amplification factor per time step.[9] For example, for the Backward/Forward scheme (2.2) we obtain

$$r(\alpha, \beta) = \frac{1+\beta}{1-\alpha}. \tag{6.10}$$

More generally, the CLM method (5.1) applied to (6.8) yields a solution of the form

$$\psi^n = \sum_{j=1}^{m} d_j [r_j(\alpha, \beta)]^n, \tag{6.11}$$

[9] The notation is inconsistent here: a superscript on ψ means a time level, whereas r is raised to the nth power.

where $r_j(\alpha, \beta)$, $j = 1, \dots, m$, are the roots of the corresponding *stability polynomial*

$$P(z; \alpha, \beta) := \sum_{j=0}^{m} (c_j - \alpha a_j - \beta b_j)\, z^{m-j} \tag{6.12}$$

and d_j is a constant (if r_j is a simple root) or a polynomial in n of degree less than q (if r_j is a root of multiplicity $q > 1$). For example, for the trapezoidal/leapfrog scheme (2.3) we have

$$r_{1,2}(\alpha, \beta) = \frac{\beta \pm \sqrt{1 + \beta^2 - \alpha^2}}{1 - \alpha}. \tag{6.13}$$

For each $j = 1, \dots, m$, $r_j(\alpha, \beta)$ is the amplification factor for the corresponding solution mode in (6.11); the solution is bounded (for all initial conditions) if and only if these roots satisfy the root condition (5.12). The stability polynomial (5.11) and characteristic polynomial (6.12) are related by $P_0(z) = P(z; 0, 0)$. Since the roots of a polynomial are continuous functions of the polynomial coefficients, each root $r_j(\alpha, \beta)$ of $P(z; \alpha, \beta)$ tends toward a root r_j of $P_0(z)$ as $\alpha, \beta \to 0$ (we will index these roots so that $r_j(0, 0) = r_j$, $j = 1, \dots, m$). Thus, absolute stability implies zero stability but not conversely.[10] Also, since $P_0(1) = 1$ for a consistent method and we have set $r_1 = 1$, we have $r_1(\alpha, \beta) \to 1$ as $\alpha, \beta \to 0$. We identify this root as the *physical mode* and the remaining roots $r_j(\alpha, \beta)$, $j = 2, \dots, m$, as *computational modes.*

The *region of absolute stability* for a semi-implicit method is the set of all values (α, β) for which $|r_j(\alpha, \beta)| < 1$ for $j = 1, \dots, m$.[11] Since in general we may regard a and b in the test equation (6.8) as complex constants, the stability region is a subset of $\mathbb{C} \times \mathbb{C}$, which (as a vector space over the reals) is four-dimensional. Thus, describing that region in general seems daunting. Rather, we can consider separately cases where α or β is purely real or imaginary and thus reduce the dimension. Ascher et al. [2] considered the case α real (negative) and β pure imaginary, corresponding to the convection-diffusion equation with the diffusion treated implicitly. Here we consider the purely oscillatory case where both α and β are pure imaginary, corresponding to the case of fast and slow waves as in the shallow-water model above. In particular, we use $\alpha = i\Omega_f$ and $\beta = i\Omega_s$, with the real parameters $\Omega_f = \omega_f \Delta t$ and $\Omega_s = \omega_s \Delta t$ being the Courant numbers for the fast components (e.g., gravity waves) and slow components (e.g., advection), respectively, of the solution. Thus, we can evaluate the absolute stability of a scheme (for oscillatory problems) by plotting the intersection of the stability region with the (Ω_f, Ω_s)-plane.

An ideal semi-implicit method for oscillatory problems would satisfy the following criteria:

1. stable for $\Omega_s = O(1)$ and for all Ω_f,
2. damps the solution for $|\Omega_f| \gg 1$,
3. has high-order accuracy.

[10] More precisely: if there is a positive time step Δt_0 and values of α and β such that a given method is absolutely stable for all Δt in $(0, \Delta t_0]$ with that α and β, then the method is zero-stable.

[11] Actually, this is the interior of the region; some points on the boundary may also be included.

Condition (1) is the best stability we can hope for (since purely explicit schemes have CFL conditions of the form $\Omega_s = O(1)$). Condition (2) is desirable, since any fast modes with large Courant numbers will be poorly approximated (although stable), and thus should be removed. Condition (3) is a major goal of this work, since high accuracy in time may be beneficial in models with accurate space discretizations, but most semi-implicit methods used to date are at best second-order accurate. In the subsections which follow we look for such schemes by considering the possible schemes in order of increasing number of steps m.

6.3 One-step semi-implicit schemes

For a one-step scheme ($m = 1$), the requirements (5.7) and (5.9) for consistency reduce the general CLM method (5.1) to

$$\frac{\psi^{n+1} - \psi^n}{\Delta t} = \theta \mathcal{A}(\psi^{n+1}) + (1-\theta)\mathcal{A}(\psi^n) + \mathcal{B}(\psi^n), \tag{6.14}$$

where θ is the single free parameter. From (5.10) the corresponding truncation error is

$$\tau = \left[\frac{1}{2}\mathcal{B}'(t_n) + \left(\frac{1}{2} - \theta\right)\mathcal{A}'(t_n)\right]\Delta t + O\left((\Delta t)^2\right). \tag{6.15}$$

Thus, this scheme is first-order accurate, unless $\theta = \frac{1}{2}$ (the trapezoidal implicit scheme) in which case it is second-order accurate in the implicit terms. The corresponding region of absolute stability can be shown to be

$$\left[\Omega_s + (1-\theta)\Omega_f\right]^2 \le \theta^2 \Omega_f^2. \tag{6.16}$$

Thus, this method is absolutely *unstable* for any $\Omega_s \neq 0$. This result for forward (Euler) time differencing is of course well-known; here, it says that there are no one-step semi-implicit methods suitable for use in oscillatory problems.

6.4 Two-step semi-implicit schemes

For a two-step scheme ($m = 2$), the requirements (5.7) and (5.8) for second-order accuracy reduce the general CLM method (5.1) to

$$\begin{aligned} \frac{(\gamma + \frac{1}{2})\psi^{n+1} - 2\gamma\psi^n + (\gamma - \frac{1}{2})\psi^{n-1}}{\Delta t} \quad = \quad & \left(\gamma + \tfrac{c}{2}\right)\mathcal{A}(\psi^{n+1}) + (1 - \gamma - c)\mathcal{A}(\psi^n) + \tfrac{c}{2}\mathcal{A}(\psi^{n-1}) \\ & + (1+\gamma)\mathcal{B}(\psi^n) - \gamma\mathcal{B}(\psi^{n-1}), \end{aligned} \tag{6.17}$$

with two parameters c and γ left to specify. Some special cases include $(\gamma, c) = (0, 1)$ for the trapezoidal/leapfrog scheme (2.3) and $(\gamma, c) = (\frac{1}{2}, 0)$ for a semi-implicit AB2 scheme. An analysis shows that:

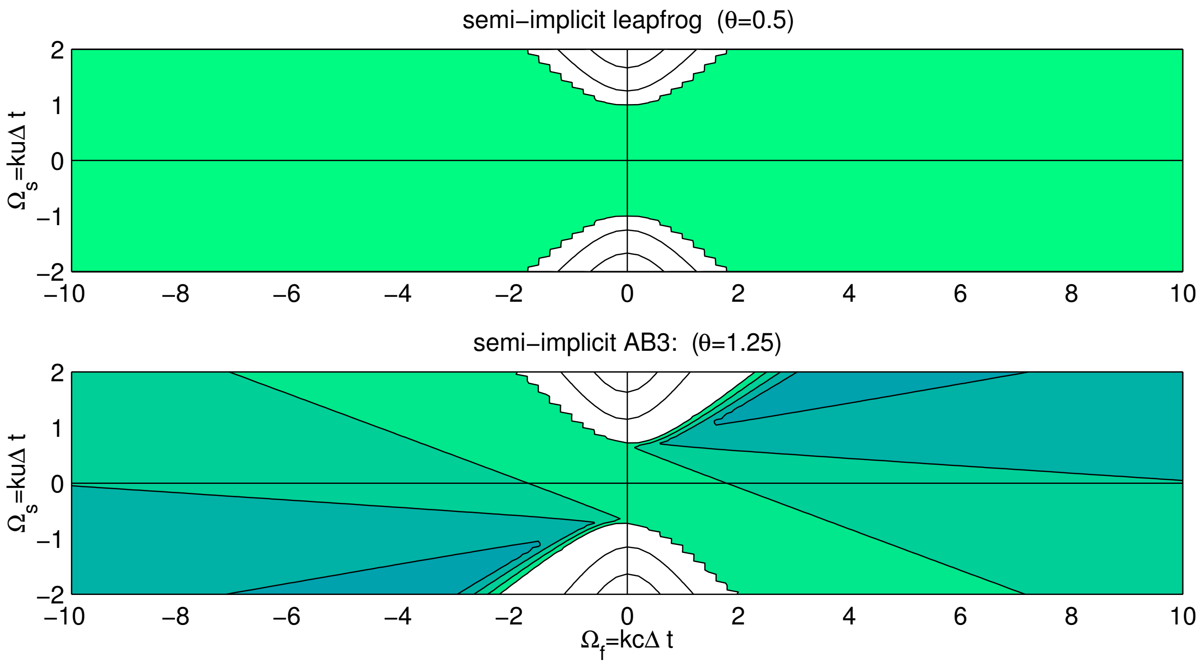

Figure 3: Amplification factors for the trapezoidal/leapfrog scheme (top panel) and the semi-implicit AB3 scheme (bottom panel) for oscillatory problems. The region of absolute stability is shaded, and the contour interval for the amplification factor is 0.1 for $|r| \leq 1$ and and 1.0 for $|r| > 1$.

- Only $\gamma = 0$ (leapfrog explicit) gives a nontrivial *explicit* stability region, i.e., stability for some $|\Omega_s| > 0$
- With $\gamma = 0$, $c \geq \frac{1}{2}$ is needed for stability as $|\Omega_f| \to \infty$. The stability region is given by $[\Omega_s + (1-c)\Omega_f]^2 \leq 1 + c^2\Omega_f^2$, which includes $|\Omega_s| \leq \sqrt{2c-1}/c$. This is maximized by choosing $c = 1$, i.e., the trapezoidal implicit scheme.

Thus, we conclude that of the family (6.17) of second-order two-step semi-implicit schemes, the traditional trapezoidal/leapfrog scheme (2.3) is best for oscillatory problems. However, note that it suffers from two defects: it has a computational mode which is not damped, and for large Ω_f the stability is only marginal, i.e., the high-frequency modes are not damped. The absolute stability region for this scheme is shown in the upper panel of Fig. 3.

6.5 Three-step semi-implicit schemes (generalized Adams form)

For a three-step scheme ($m = 3$), there are many possibilities. Here we consider only schemes with $c_2 = c_3 = 0$, which we refer to as the *generalized Adams form* (since it generalizes the explicit Adams-Bashforth and implicit Adams-Moulton methods). These schemes have two distinct advantages: their computational modes are well-behaved, damping as $\Delta t \to 0$ (the stability polynomial has a single nonzero root $r_1 = 1$), and they do not require storing past values of ψ. Such schemes take the form

$$\begin{aligned}\frac{\psi^{n+1} - \psi^n}{\Delta t} &= a_0\mathcal{A}(\psi^{n+1}) + a_1\mathcal{A}(\psi^n) + a_2\mathcal{A}(\psi^{n-1}) + a_3\mathcal{A}(\psi^{n-2}) \\ &\qquad + b_1\mathcal{B}(\psi^n) + b_2\mathcal{B}(\psi^{n-1}) + b_3\mathcal{B}(\psi^{n-2}), \end{aligned} \tag{6.18}$$

Requiring third-order accuracy for the explicit part gives the third-order Adams-Bashforth method (AB3), i.e.,

$$b_1 = \tfrac{23}{12}, \qquad b_2 = -\tfrac{16}{12}, \qquad b_3 = \tfrac{5}{12}. \tag{6.19}$$

If we set $a_3 = 0$ for simplicity, we can require second-order accuracy for the implicit part, yielding the generalized second-order Adams-Moulton method (GAM2)

$$a_0 = \theta, \qquad a_1 = \tfrac{3}{2} - 2\theta, \qquad a_2 = -\tfrac{1}{2} + \theta, \tag{6.20}$$

where θ is a free parameter. Several choices of θ are of interest. The choice $\theta = 0$ gives the (explicit) AB2 scheme. The choice $\theta = \frac{1}{2}$ gives the trapezoidal implicit scheme, which omits ψ^{n-1}. The choice $\theta = \frac{5}{12}$ minimizes the truncation error, giving the classical third-order Adams-Moulton (AM3) method. We will refer to the scheme (6.18) with the coefficients (6.19) and (6.20) as the SI2/AB3 scheme; this scheme does not appear to have been studied before. Analysis of this scheme shows:

- Stability for $|\Omega_f| \to \infty$ (with $|\Omega_s| = 0$) requires $\theta \geq \frac{1}{2}$
- For $\theta \geq \frac{9}{16}$, high frequencies damp with the factor $\lambda = (\theta - \frac{1}{2})/\theta$
- For $\Omega_s \neq 0$, the trapezoidal ($\theta = \frac{1}{2}$) and AM3 ($\theta = \frac{5}{12}$) versions are *unstable*
- For $\theta \approx \frac{5}{4}$, the SI2/AB3 scheme is *stable*

The stability region for the SI2/AB3 scheme with various choices of the free parameter θ are shown in Fig. 4. From these results it would appear that the optimum stability properties are obtained when $\theta \approx 1.25$. For comparison with the trapezoidal/leapfrog scheme, the corresponding absolute stability region is is shown in the lower panel of Fig. 3.

If we allow $a_3 \neq 0$, we can require third-order accuracy, yielding the generalized third-order Adams-Moulton method (GAM3)

$$a_0 = \theta, \qquad a_1 = \tfrac{23}{12} - 3\theta, \qquad a_2 = -\tfrac{16}{12} + 3\theta, \qquad a_3 = \tfrac{5}{12} - \theta, \tag{6.21}$$

where θ is a free parameter. Several choices of θ are of interest. The choice $\theta = 0$ gives the (explicit) AB3 scheme. The choice $\theta = \frac{5}{12}$ gives the classical third-order Adams-Moulton (AM3), which omits ψ^{n-2}. The choice $\theta = \frac{3}{8}$ minimizes the truncation error, giving the classical fourth-order Adams-Moulton (AM4) method. We will refer to the scheme (6.18) with the coefficients (6.19) and (6.21) as the SI3/AB3 scheme. Preliminary analysis of this scheme suggests that there is *no* value of θ which yields adequate absolute stability for oscillatory problems.

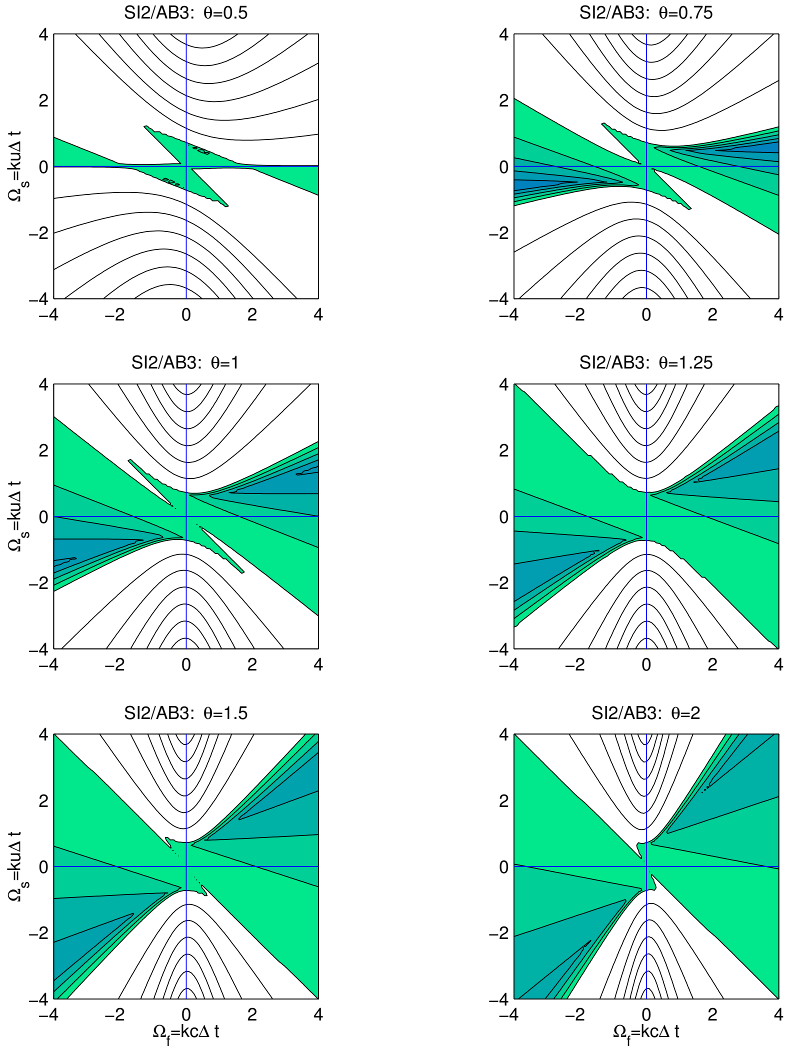

Figure 4: Amplification factors for the semi-implicit AB3 scheme for oscillatory problems. The six panels correspond to the values $\theta = 0.5$, 0.75, 1.0, 1.25, 1.5, and 2 as labeled. Contours and shading are as in Fig. 3.

7 Conclusions

We have reviewed the formulation and analysis of semi-implicit time differencing schemes for partial differential equations. In general, combining explicit and implicit schemes which are both convergent leads to a convergent method. However, we have shown that the resulting implicit equations must be solved as an implicit problem (either directly or by an iterative method which couples the space and time discretizations). If this solution is avoided by using simple fixed-point iteration instead, the iteration will converge only when the time step is small, thus defeating the purpose of semi-implicit time differencing.

A search through part of the vast space of semi-implicit schemes (of the CLM type) has yielded the following results for purely oscillatory problems of the type frequently encountered in atmospheric modeling:

- There are no useful one-step semi-implicit schemes.
- The best two-step semi-implicit scheme is the widely-used trapezoidal/leapfrog scheme; however, this includes a non-damped computational mode, does not damp the poorly-approximated high-frequency modes, and is only second-order accurate.
- The semi-implicit second-order Adams-Bashforth (AB2) scheme used by Simmons and Temperton [16] is unstable.
- The combination of the backward scheme with the third-order Adams-Bashforth (AB3) scheme (proposed for use in the CSU GCM) is unstable. *Details to be included in the next draft.*
- The combination of the trapezoidal scheme with the third-order Adams-Bashforth (AB3) scheme is unstable, as concluded by Durran [7]
- There exists a stable three-step scheme which combines the AB3 explicit scheme with a generalized second-order Adams-Moulton implicit scheme.

It appears that the SI2/AB3 scheme identified here is a promising candidate for accurate integration of atmospheric models. Future work will include testing this scheme in shallow-water and other models and continuing the search for other high-order semi-implicit schemes.

References

[1] Ascher, U. M., and L. R. Petzold, 1998: *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations.* Society for Industrial and Applied Mathematics, 314 pp.

[2] Ascher, U. M., S. J. Ruuth, and B. T. R. Wetton, 1995: Implicit-explicit methods for time-dependent PDE's. *SIAM J. Numer. Anal.*, **32**, 797–823.

[3] Atkinson, K. E., 1989: *An Introduction to Numerical Analysis*, second edition. Wiley, 712 pp.

[4] Cooper, G. J., and A. Sayfy, 1980: Additive methods for the numerical solution of ordinary differential equations. *Math. Comput.*, **35**, 1980, 1159–1172.

[5] Cooper, G. J., and A. Sayfy, 1983: Additive Runge-Kutta methods for stiff ordinary differential equations. *Math. Comput.*, **40**, 1983, 207–218.

[6] Côté, J., M. Béland, and A. Staniforth, 1983: Stability of vertical discretization schemes for semi-implicit primitive equation models: Theory and application. *Mon. Wea. Rev.*, **111**, 1189–1207.

[7] Durran, D. R., 1991: The third-order Adams-Bashforth method: An attractive alternative to leapfrog time differencing. *Mon. Wea. Rev.*, **119**, 702–720.

[8] Fulton, S. R., 1993: A semi-implicit spectral method for the anelastic equations. *J. Comp. Phys.*, **106**, 299–305.

[9] Heikes, R., and D. A. Randall, 1995: Numerical integration of the shallow-water equations on a twisted icosahedral grid. Part I: Basic design and results of tests. *Mon. Wea. Rev.*, **123**, 1862–1880.

[10] Lambert, J. D., 1991: *Numerical Methods for Ordinary Differential Systems.* Wiley, 293 pp.

[11] Kang, Y., 1994: Analysis and Applications of Combined Numerical Methods for Systems of Differential Equations. Ph.D. Dissertation, Department of Mathematics and Computer Science, Clarkson University, 157 pp.

[12] Kwizak, M. and A. J. Robert, 1971: A semi-implicit scheme for gridpoint atmospheric models of the primitive equations. *Mon. Wea. Rev.*, **99**, 32–36.

[13] Robert, A. J., J. Henderson, and C. Turnbull, 1972: An implicit time integration scheme for baroclinic modes of the atmosphere. *Mon. Wea. Rev.*, **100**, 329–335.

[14] Sadun, L., 2001: *Applied Linear Algebra.* Prentice-Hall, 349 pp.

[15] Simmons, A. J., B. J. Hoskins, and D. M. Burridge, 1978: Stability of the semi-implicit method of time integration. *Mon. Wea. Rev.*, **106**, 405–412.

[16] Simmons, A. J., and C. Temperton, 1997: Stability of a two-time-level semi-implicit integration scheme for gravity wave motion. *Mon. Wea. Rev.*, **125**, 600–615.

[17] Staniforth, A., and R. W. Daley, 1979: A baroclinic finite-element model for regional forecasting with the primitive equations. *Mon. Wea. Rev.*, **107**, 107–121.

[18] Stoer, J., and R. Bulirsch, 1992: *Introduction to Numerical Analysis*, second edition. Springer, 660 pp.