Semi-Implicit Discretization of the Shallow-Water Equations

Scott R. Fulton

15 June 2007

Abstract

The shallow-water equations admit both fast (gravity-inertia) and slow (Rossby) modes. In most situations of meteorological and oceanographic interest the fast modes contain little energy. Therefore, semi-implicit time discretization has the potential to increase the time step beyond the explicit CFL limit without compromising accuracy. These notes analyze the linearized shallow-water equations to identify which terms should be treated implicitly and explicitly. Based on this analysis, appropriate splittings are developed for various forms of the nonlinear shallow-water equations. Finally, the properties of various discretizations are surveyed to motivate the choice of a good semi-implicit scheme.

1 Introduction

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

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} + f\mathbf{k} \times \mathbf{v} + g\nabla h = \mathbf{F}, \tag{1.1}$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (h\mathbf{v}) = F_h, \tag{1.2}$$

where $\mathbf{v}$ is the (horizontal) vector velocity, 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 F_h for momentum and mass are regarded as known functions of t (and possibly of $\mathbf{v}$ and h). Other forms of these equations will also be considered below.

It is well-known that the shallow-water equations admit both fast (gravity-inertia) and slow (Rossby) modes. In most situations of meteorological and oceanographic interest, only the slow modes are of interest, and the fast modes contain little energy. Therefore, a natural (and common) solution procedure is to split the terms in the equations into two groups, namely, those related to the fast and slow modes, and treat the former implicitly and the latter explicitly. Such a *splitting* can be written generically in the form

$$\psi'(t) = \mathcal{A}(\psi(t)) + \mathcal{B}(\psi(t)), \tag{1.3}$$

where the vector ψ represents the solution at time t and the operators $\mathcal{A}$ and $\mathcal{B}$ represent the terms to be treated implicitly and explicitly, respectively. For example, if the space dependence is left continuous, we can set $\psi = [u, v, h]^T$, where u and v are components of the vector velocity. Likewise, if the equations are discretized in space (e.g., by a finite-difference or spectral method), we can set $\psi = [\mathbf{u}^T, \mathbf{v}^T, \mathbf{h}^T]^T$, where the components are (mathematical) vectors (e.g., collections of gridpoint values or spectral coefficients) representing u, v, and h at time t. Similar interpretations of ψ are possible for other forms of the shallow-water equations (see below).

We consider discretizations of (1.3) which replace the solution $\psi(t)$ with an approximation $\psi^n \approx \psi(t_n)$ at times $t_n = n\Delta t$ for $n = 0, 1, \ldots$. 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 $\mathcal{B}(\psi) \equiv 0$. The simplest such scheme combines the backward (implicit Euler) and forward (explicit Euler) schemes to give the discretization

$$\frac{\psi^{n+1} - \psi^n}{\Delta t} = \mathcal{A}(\psi^{n+1}) + \mathcal{B}(\psi^n). \tag{1.4}$$

The scheme most often used in meteorological models (e.g., [10, 3, 12, 14]) combines the trapezoidal (implicit) and leapfrog (explicit) schemes, yielding the discretization

$$\frac{\psi^{n+1} - \psi^{n-1}}{2\Delta t} = \frac{\mathcal{A}(\psi^{n-1}) + \mathcal{A}(\psi^{n+1})}{2} + \mathcal{B}(\psi^n). \tag{1.5}$$

As we will see in section 4 below, more general combinations of linear multistep methods are possible.[1] Any such scheme results in an implicit problem of the form

$$\psi^{n+1} - \tau\mathcal{A}(\psi^{n+1}) = \mathcal{C}^n, \tag{1.6}$$

where τ is a multiple of the time step ($\tau = \Delta t$ for the two schemes above) and $\mathcal{C}^n$ is a combination of values of $\mathcal{A}(\psi)$ and $\mathcal{B}(\psi)$ which are known at time t_n. The execution of one time step thus consists of first computing $\mathcal{C}^n$ and then solving (1.6)—with appropriate boundary conditions—for ψ^{n+1}.

Formulating a good semi-implicit scheme for the shallow-water equations thus involves two choices:

- *Splitting*: which terms should be treated implicitly and explictly?
- *Discretization*: which implicit and explicit disretizations should be used?

The purpose of these notes is to answer these questions. First, the choice of splitting involves identifying the terms responsible for the fast modes; this we do in section 2 by analyzing the linearized shallow-water equations. The choice of splitting also affects the complexity of the resulting implicit problem (1.6); the forms this takes for various forms of the shallow-water model are derived in section 3. Section 4 defines general semi-implicit schemes based on linear multistep methods and summarizes conditions for their consistency and convergence. Absolute stability is defined and analyzed in section 5, and a particularly promising semi-implicit scheme is identified. Section 6 summaries the conclusions.

[1] It is also possible to formulate semi-implicit schemes based on Runge-Kutta methods. For example, combining the backward and fourth-order Runge-Kutta methods generates a scheme which works well for problems of the advection-diffusion type [5]. However, an extensive search [9] suggests that no such schemes exist with stability properties appropriate for oscillatory problems like the shallow-water equations; see also section 5.

2 Linear analysis

To justify a proper splitting of terms for semi-implicit time differencing, we use the following linear analysis.[2] Assuming that f is constant and the forcing vanishes, and using Cartesian coordinates with velocity components u and v in the x and y directions, we can linearize (1.1)–(1.2) about a constant basic state with velocity $(\bar{u}, \bar{v})$ and height $\bar{h} > 0$ to obtain

$$u_t + \bar{u}u_x + \bar{v}u_y - fv + gh_x = 0, \tag{2.1}$$

$$v_t + \bar{u}v_x + \bar{v}v_y + fu + gh_y = 0, \tag{2.2}$$

$$h_t + \bar{u}h_x + \bar{v}h_y + \bar{h}(u_x + v_y) = 0. \tag{2.3}$$

Here, h now denotes the *deviation* of the free surface height from the reference height $\bar{h}$, and subscripts denote partial derivatives. Scaling h by g/c, where $c := \left(g\bar{h}\right)^{1/2}$ is the phase speed of a pure gravity wave, we can write (2.1)–(2.3) in the symmetrized matrix form

$$\frac{\partial \psi}{\partial t} = \mathcal{L}\psi, \tag{2.4}$$

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

$$\mathcal{L} = \begin{bmatrix} -\bar{u}\partial_x - \bar{v}\partial_y & f & -c\partial_x \\ -f & -\bar{u}\partial_x - \bar{v}\partial_y & -c\partial_y \\ -c\partial_x & -c\partial_y & -\bar{u}\partial_x - \bar{v}\partial_y \end{bmatrix}. \tag{2.5}$$

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

$$\frac{d\hat{\psi}}{dt} = L\hat{\psi}, \tag{2.6}$$

where

$$L = \begin{bmatrix} -i\bar{\omega} & f & -ikc \\ -f & -i\bar{\omega} & -ilc \\ -ikc & -ilc & -i\bar{\omega} \end{bmatrix} \tag{2.7}$$

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

$$\psi(x, y, t) = \Psi_j e^{i(kx+ly-\omega_j t)}. \tag{2.8}$$

[2] A linear analysis based on the vorticity/divergence form yields the same conclusions.

Direct calculation yields the following eigenvalues and eigenvectors:

$$\text{Slow (geostrophic) modes:} \qquad \omega_1 = \bar{\omega}, \qquad \Psi_1 = \frac{1}{\nu}\begin{bmatrix} -ilc \\ ikc \\ f \end{bmatrix}, \tag{2.9}$$

$$\text{Fast (gravity-inertia) modes:} \quad \omega_{2,3} = \bar{\omega} \pm \nu, \qquad \Psi_{2,3} = \frac{1}{\nu\bar{k}\sqrt{2}}\begin{bmatrix} \pm\nu k + ilf \\ \pm\nu l - ikf \\ c\bar{k}^2 \end{bmatrix}, \tag{2.10}$$

where $\nu := (f^2 + \omega_g^2)^{1/2}$, $\omega_g = \bar{k}c$, and $\bar{k} := (k^2 + l^2)^{1/2}$. Since L is skew-Hermitian, the eigenvectors Ψ_j are orthogonal in the standard inner product $\langle \Psi, \Phi \rangle := \Psi^* \Phi$ (where the asterisk denotes the conjugate transpose); they have been scaled so that they have unit norm in the corresponding norm $\|\Psi\| := \sqrt{\langle \Psi, \Psi \rangle}$.

To determine an appropriate splitting of terms for semi-implicit time differencing, we write (2.6) in the form

$$\frac{d\hat{\psi}}{dt} = A\hat{\psi} + B\hat{\psi} + C\hat{\psi}, \tag{2.11}$$

where

$$A = \begin{bmatrix} 0 & 0 & -ikc \\ 0 & 0 & -ilc \\ -ikc & -ilc & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -i\bar{\omega} & 0 & 0 \\ 0 & -i\bar{\omega} & 0 \\ 0 & 0 & -i\bar{\omega} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & f & 0 \\ -f & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.12}$$

The matrix A corresponds to the height-gradient term $g\nabla h$ in the momentum equation and the divergence term $\bar{h}\nabla \cdot \mathbf{v}$ in the continuity equation, which we identify as the *gravity-wave terms* as explained below; the matrices B and C correspond to the advective and and Coriolis terms, respectively. The corresponding spectral radii $\rho(A) = \omega_g$, $\rho(B) = \bar{\omega}$, and $\rho(C) = f$ typically satisfy

$$\omega_g \gg \bar{\omega} \gg f, \tag{2.13}$$

corresponding to a separation of time scales. In particular, the ratio $\rho(B)/\rho(A) = \bar{\omega}/\omega_g = \bar{u}/c$ is the Froude number, which is typically much smaller than one. For example, using the values $c \sim 300\,\mathrm{ms}^{-1}$, $\bar{u} \sim 30\,\mathrm{ms}^{-1}$, $f \sim 10^{-4}\,\mathrm{s}^{-1}$, and $\bar{k} \sim 2\pi/(200\,\mathrm{km})$ leads to $\omega_g/\bar{\omega} \sim 10$ and $\bar{\omega}/f \sim 10$. Thus, we can think of the terms A, B, and C in (2.11) as the "fast", "slower", and "slowest" terms, respectively.

More precisely, we can quantify which of terms A, B, and C are important for which types of motion by computing the size of each term applied to each of the (normalized) modes Ψ_j ($j = 1, 2, 3$). For the slow (geostrophic) modes ($j = 1$) we obtain

$$\|A\Psi_1\| = \frac{f\omega_g}{\nu} \approx f, \qquad \|B\Psi_1\| = \bar{\omega}, \qquad \|C\Psi_1\| = \frac{f\omega_g}{\nu} \approx f, \tag{2.14}$$

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

$$\|A\Psi_{2,3}\| = \frac{\omega_g}{\nu}\left(\frac{\omega_g^2+\nu^2}{2}\right)^{1/2} \approx \omega_g, \quad \|B\Psi_{2,3}\| = \bar{\omega}, \quad \|C\Psi_{2,3}\| = \frac{f}{\nu}\left(\frac{f^2+\nu^2}{2}\right)^{1/2} \approx \frac{f}{\sqrt{2}}. \tag{2.15}$$

Using the scales assumed above we thus find that the Coriolis terms (C) have (approximate) size f and the advection terms (B) have size $10f$ for both the slow and fast modes, while the gravity-wave terms (A) have (approximate) size f for the slow modes but $100f$ 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 if and only if fast modes contain little energy; in this situation the advective terms can 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? Arguments can be made for both approaches. Reasons for treating the Coriolis terms explicitly include:

- The time scale is slower than for advection, so an implicit method is not needed for stability.
- The resulting implicit system will be simpler.

Reasons for treating the Coriolis terms implicitly include:

- Implicit methods generally have smaller truncation error than explicit methods (of the same order), so the accuracy may be higher.
- Treating the Coriolis and height gradient terms with the same method may give a better representation of geostrophic balance.
- The operators $A + C$ and B of the associated splitting commute and thus are orthogonally diagonalizable, so the stability analysis of section 5 applies directly (to the linear problem), giving some confidence that methods which the analysis identifies as good will in fact work properly.

In the next section we detail both of these approaches for the nonlinear shallow-water equations in various forms.

3 Splitting the nonlinear shallow-water equations

We now seek appropriate splittings for the nonlinear shallow-water equations based on the above analysis. To treat the gravity-wave terms implicitly, we specify a positive reference height $\bar{h}$ and set $h = \bar{h} + h'$; while $\bar{h}$ must be constant in time it may vary in space (we also consider the simplifications possible when $\bar{h}$ is is constant in space). In the sections which follow we develop splittings for the equations in momentum form and vorticity/divergence form; for each we show how to treat the Coriolis terms either explicitly or implicitly.

3.1 Momentum form

In the momentum form (1.1)–(1.2) of the shallow-water equations the predictive variables are $\mathbf{v}$ and h. While the momentum equation is normally written in the advective form (1.1), in some models (e.g., [6]) it is written in the equivalent rotational form

$$\frac{\partial \mathbf{v}}{\partial t} + (f + \zeta)\mathbf{k} \times \mathbf{v} + \nabla(gh + K) = \mathbf{F}, \tag{3.1}$$

where $\zeta = \mathbf{k} \cdot \nabla \times \mathbf{v}$ is the relative vorticity and $K = \frac{1}{2}\mathbf{v} \cdot \mathbf{v}$ is the specific kinetic energy. Since only linear terms will be treated implicitly and these are identical in both the advective and rotational forms, both forms of the momentum equation will lead to the same semi-implicit discretizations. Therefore, we consider only the momentum form here.

3.1.1 Coriolis terms treated explicitly

Substituting $h = \bar{h} + h'$ into (1.1)–(1.2) and moving the terms to be treated explicitly to the right-hand side yields

$$\frac{\partial \mathbf{v}}{\partial t} + g\nabla h \qquad = \mathbf{F} - \mathbf{v} \cdot \nabla \mathbf{v} - f\mathbf{k} \times \mathbf{v}, \tag{3.2}$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (\bar{h}\mathbf{v}) = F_h - \nabla \cdot (h'\mathbf{v}). \tag{3.3}$$

Comparing (3.2)–(3.3) with (1.3), we see that $\mathcal{B}$ represents all terms on the right, while $-\mathcal{A}$ represents the terms (other than the time derivatives) on the left. Therefore, for any semi-implicit scheme the implicit problem (1.6) corresponding to (3.2)–(3.3) takes the form

$$\mathbf{v} + \tau g \nabla h \qquad = \mathbf{V}, \tag{3.4}$$

$$h + \tau \nabla \cdot (\bar{h}\mathbf{v}) = H, \tag{3.5}$$

where $\mathbf{v}$ and h now represent the variables *at the new time level* (i.e., ψ^{n+1} with the superscript dropped for simplicity), and $\mathbf{V}$ and H are computed explicitly from values at the old time level (i.e., $\mathcal{C}^n$ with the superscript dropped for simplicity).

This system is easily solved by eliminating $\mathbf{v}$ to obtain the linear elliptic problem

$$h - \tau^2 \nabla \cdot \left(\bar{c}^2 \nabla h\right) = G, \tag{3.6}$$

where $\bar{c} = \left(g\bar{h}\right)^{1/2}$ and

$$G := H - \tau \nabla \cdot \left(\bar{h}\mathbf{V}\right). \tag{3.7}$$

If $\bar{h}$ is constant in space, then (3.6) further reduces to the modified Helmholtz equation

$$h - \tau^2 \bar{c}^2 \nabla^2 h = G. \tag{3.8}$$

Solving (3.8) straightforward (with any space discretization), and the slight generalization (3.6) is not much more difficult in cases where multigrid methods can be used. Once h is obtained by solving (3.6) or (3.8), $\mathbf{v}$ can be obtained from (3.4).

3.1.2 Coriolis terms treated implicitly

If we treat the Coriolis terms implicitly, then (3.2) is replaced by

$$\frac{\partial \mathbf{v}}{\partial t} + f\mathbf{k} \times \mathbf{v} + g\nabla h = \mathbf{F} - \mathbf{v} \cdot \nabla \mathbf{v}, \tag{3.9}$$

where the terms to be treated explicitly are on the right-hand side. The implicit problem corresponding to (3.9) and (3.3) is

$$\mathbf{v} + \tau f \mathbf{k} \times \mathbf{v} + \tau g \nabla h = \mathbf{V}, \tag{3.10}$$

$$h + \tau \nabla \cdot (\bar{h}\mathbf{v}) \qquad = H, \tag{3.11}$$

where $\mathbf{V}$ is now different than in (3.4). This system also can be solved by eliminating $\mathbf{v}$ as follows. Taking the cross product of $\tau f \mathbf{k}$ with (3.10) and subtracting this from (3.10) yields

$$\mathbf{v} = a\left[(\mathbf{V} - \tau f \mathbf{k} \times \mathbf{V}) - \tau g \left(\nabla h - \tau f \mathbf{k} \times \nabla h\right)\right] \tag{3.12}$$

where $a = \left(1 + \tau^2 f^2\right)^{-1}$ (usually $|\tau f| \ll 1$ so $a \approx 1$). Substituting this into (3.11) yields

$$h + \tau^2 \mathcal{J}\left(h, \tau f \bar{c}^2\right) - \tau^2 \nabla \cdot \left(\bar{c}^2 \nabla h\right) = G, \tag{3.13}$$

where now $\bar{c} = \left(ag\bar{h}\right)^{1/2}$, $\mathcal{J}(\alpha, \beta) = \mathbf{k} \cdot (\nabla\alpha \times \nabla\beta)$ is the Jacobian operator, and

$$G = H - \tau \nabla \cdot \left[a\bar{h}\left(\mathbf{V} - \tau f \mathbf{k} \times \mathbf{V}\right)\right]. \tag{3.14}$$

Since $\bar{c}$ and f are known and $\bar{c} > 0$, (3.13) is a linear elliptic equation for h. It may also be written in the form

$$h - \tau^2 \mathbf{b} \cdot \nabla h - \tau^2 \bar{c}^2 \nabla^2 h = G \tag{3.15}$$

where

$$\mathbf{b} = \nabla\left(\bar{c}^2\right) + \tau \mathbf{k} \times \nabla\left(f\bar{c}^2\right). \tag{3.16}$$

If $\bar{h}$ does not vary too rapidly in space, the first-order term in (3.13) [or (3.15)] will be relatively small so the second-order term will dominate; thus we should be able to solve this equation for h using standard methods. The corresponding velocity $\mathbf{v}$ then can be computed from (3.12). Note that if f is constant (i.e., an f-plane) and $\bar{h}$ is constant, then $\mathbf{b} = 0$ and (3.15) reduces to the modified Helmholtz equation (3.8), but this time with G given by (3.14).

3.2 Vorticity/divergence form

The momentum form (1.1)–(1.2) 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, some authors advocate using vorticity and divergence instead (e.g., [7]). To do so, we take the dot product of $\mathbf{k}$ with the curl of the momentum equation in the form (3.1) to obtain the vorticity equation

$$\frac{\partial \eta}{\partial t} + \nabla \cdot (\eta \mathbf{v}) = F_\eta := \mathbf{k} \cdot \nabla \times \mathbf{F}. \tag{3.17}$$

where $\eta := f + \zeta$ is the absolute vorticity. Likewise, taking the divergence of (3.1) gives the divergence equation

$$\frac{\partial \delta}{\partial t} - \mathbf{k} \cdot \nabla \times (\eta \mathbf{v}) + \nabla^2 (gh + K) = F_\delta := \nabla \cdot \mathbf{F}, \tag{3.18}$$

where $\delta = \nabla \cdot \mathbf{v}$ is the divergence. Combining these equations with (1.2) gives a system for predicting η (or ζ), δ, and h. To close this system we introduce a velocity potential χ and streamfunction ψ satisfying

$$\mathbf{v} = \nabla \chi + \mathbf{k} \times \nabla \psi. \tag{3.19}$$

Then the velocity $\mathbf{v}$ can be obtained from ζ and δ by solving the Poisson problems

$$\nabla^2 \psi = \zeta, \qquad \nabla^2 \chi = \delta \tag{3.20}$$

(with appropriate boundary conditions) and using (3.19).

3.2.1 Coriolis terms treated explicitly

Substituting $h = \bar{h} + h'$ into (3.17)–(3.18) and moving the terms to be treated explicitly to the right-hand side yields

$$\frac{\partial \zeta}{\partial t} = F_\eta - \nabla \cdot (\eta \mathbf{v}) \tag{3.21}$$

and

$$\frac{\partial \delta}{\partial t} + g \nabla^2 h = F_\delta + \mathbf{k} \cdot \nabla \times (\eta \mathbf{v}) - \nabla^2 K. \tag{3.22}$$

Combining these equations with (3.3) and treating the terms on the left implicitly and those on the right explicitly, the corresponding implicit problem [cf. (1.6)] takes the form

$$\zeta = Z, \tag{3.23}$$

$$\delta + \tau g \nabla^2 h = D, \tag{3.24}$$

$$h + \tau \nabla \cdot (\bar{h} \mathbf{v}) = H, \tag{3.25}$$

where ζ, δ, h, and $\mathbf{v}$ are values at the new time level and related by (3.19) and (3.20), and Z, D, and H are computed explicitly from values at the old time level.

To solve this system, first note that (3.23) gives ζ explicitly, so we can determine ψ by solving (3.20). If $\bar{h}$ is constant then (3.25) reduces to

$$h + \tau\bar{h}\delta = H. \tag{3.26}$$

Eliminating δ between (3.26) and (3.24) yields the Helmholtz problem

$$h - \tau^2\bar{c}^2\nabla^2 h = G, \tag{3.27}$$

[cf. (3.8)], where $\bar{c} = \left(g\bar{h}\right)^{1/2}$ and

$$G = H - \tau\bar{h}D. \tag{3.28}$$

Once (3.27) is solved for h, the corresponding δ can be obtained from (3.26), then χ from (3.20), and $\mathbf{v}$ from (3.19). Even if $\bar{h}$ is not constant, we can still eliminate δ by first introducing

$$\hat{\chi} := \chi + \tau g h \tag{3.29}$$

to write (3.24) as

$$\nabla^2\hat{\chi} = D. \tag{3.30}$$

Assuming we can solve this (with appropriate boundary conditions) for $\hat{\chi}$, we can then substitute for $\mathbf{v}$ in (3.25) from (3.19) and eliminate χ using (3.29) to obtain the linear elliptic equation

$$h - \tau^2\nabla\cdot\left(\bar{c}^2\nabla h\right) = G \tag{3.31}$$

[cf. (3.6)], where

$$G = H - \tau\nabla\cdot(\bar{h}\nabla\hat{\chi}) + \tau\mathcal{J}\left(\bar{h},\psi\right). \tag{3.32}$$

Once (3.31) is solved for h, we can compute χ from (3.29), δ from (3.20), and $\mathbf{v}$ from (3.19).

3.2.2 Coriolis terms treated implicitly

If we treat the Coriolis terms implicitly, then (3.21) and (3.22) are replaced by

$$\frac{\partial\zeta}{\partial t} + \nabla\cdot(f\mathbf{v}) = F_\eta - \nabla\cdot(\zeta\mathbf{v}) \tag{3.33}$$

and

$$\frac{\partial\delta}{\partial t} - \mathbf{k}\cdot\nabla\times(f\mathbf{v}) + g\nabla^2 h = F_\delta + \mathbf{k}\cdot\nabla\times(\zeta\mathbf{v}) - \nabla^2 K. \tag{3.34}$$

Combining these equations with (3.3) and treating the terms on the left implicitly and those on the right explicitly, the corresponding implicit problem [cf. (1.6)] takes the form

$$\zeta + \tau\nabla\cdot(f\mathbf{v}) = Z, \tag{3.35}$$

$$\delta - \tau\mathbf{k}\cdot\nabla\times(f\mathbf{v}) + \tau g\nabla^2 h = D, \tag{3.36}$$

$$h + \tau\nabla\cdot(\bar{h}\mathbf{v}) = H, \tag{3.37}$$

where again ζ, δ, h, and $\mathbf{v}$ are values at the new time level and related by (3.19) and (3.20), and Z, D, and H are computed explicitly from values at the old time level.

In the case of an f-plane (i.e., f is constant) with $\bar{h}$ constant also, (3.35)–(3.37) reduce to

$$\zeta + \tau f \delta = Z, \tag{3.38}$$

$$\delta - \tau f \zeta + \tau g \nabla^2 h = D, \tag{3.39}$$

$$h + \tau \bar{h} \delta = H. \tag{3.40}$$

Eliminating ζ between (3.38) and (3.39) yields

$$\delta + \tau a g \nabla^2 h = a(D + \tau f Z) \tag{3.41}$$

where $a = (1 + \tau^2 f^2)^{-1}$ as before. Then eliminating δ between (3.40) and (3.41) leads to the modified Helmholtz problem

$$h - \tau^2 \bar{c}^2 \nabla^2 h = G \tag{3.42}$$

[cf. (3.8)], where $\bar{c} = \left(a g \bar{h}\right)^{1/2}$ as before and

$$G = H - \tau a \bar{h}(D + \tau f Z). \tag{3.43}$$

Once (3.42) is solved for h we can compute δ from (3.40) and ζ from (3.38), ψ and χ from (3.20), and $\mathbf{v}$ from (3.19).

Finally, with no simplifying assumptions on $\bar{h}$ and f, we can in principle solve the implicit system (3.35)–(3.37) coupled with (3.19) and (3.20) by essentially converting it to the momentum form as follows. The key is to introduce the "forced velocity" $\mathbf{V}$ computed from Z and D as follows. First, solve

$$\nabla^2 \Psi = Z, \qquad \nabla^2 X = D \tag{3.44}$$

(with appropriate boundary conditions) for Ψ and X. Then define

$$\mathbf{V} := \nabla X + \mathbf{k} \times \nabla \Psi, \tag{3.45}$$

so that $\mathbf{k} \cdot \nabla \times \mathbf{V} = Z$ and $\nabla \cdot \mathbf{V} = D$. Then the velocity $\mathbf{v}$ derived from ψ, χ, and h satisfying the implicit system (3.35)–(3.37) satisfies (3.10) and (3.11), so we can solve (3.13) for h and compute $\mathbf{v}$ from (3.12), as explained for the momentum form above. From $\mathbf{v}$ we can then compute $\delta = \nabla \cdot \mathbf{v}$ and $\zeta = \mathbf{k} \times \nabla \mathbf{v}$ and obtain ψ and χ by solving the Poisson problems (3.20).

3.3 Streamfunction/velocity potential form

While the vorticity/divergence form considered above predicts vorticity and divergence, it still explicitly involves the velocity $\mathbf{v}$. This may be removed from the formulation in favor of χ and ψ by using (3.19) and the identities

$$\nabla \cdot (\phi \mathbf{v}) = \nabla \cdot (\phi \nabla \chi) - \mathcal{J}(\phi, \psi) \tag{3.46}$$

and

$$\mathbf{k} \cdot \nabla \times (\phi \mathbf{v}) = \nabla \cdot (\phi \nabla \psi) + \mathcal{J}(\phi, \chi) \tag{3.47}$$

(where ϕ represents any scalar) and the formula

$$K = \frac{1}{2}\left[\nabla \cdot (\chi \nabla \chi) - \chi \nabla^2 \chi + \nabla \cdot (\psi \nabla \psi) - \psi \nabla^2 \psi\right] + \mathcal{J}(\psi, \chi). \tag{3.48}$$

The resulting equations are

$$\frac{\partial \eta}{\partial t} + \nabla \cdot (\eta \nabla \chi) - \mathcal{J}(\eta, \psi) = F_\eta, \tag{3.49}$$

$$\frac{\partial \delta}{\partial t} - \nabla \cdot (\eta \nabla \psi) - \mathcal{J}(\eta, \chi) + \nabla^2 (gh + K) = F_\delta, \tag{3.50}$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (h \nabla \chi) - \mathcal{J}(h, \psi) = F_h. \tag{3.51}$$

In this way the only operators needed are the flux divergence, Jacobian, and Laplacian. Since this form of the equations separates the divergent and rotational parts of the velocity, it is possible to treat only the former implicitly, which results in slightly simpler formulations as follows.

3.3.1 Coriolis terms treated explicitly

Substituting $h = \bar{h} + h'$ into (3.49)–(3.51) and moving the terms to be treated explicitly (all nonlinear and Coriolis terms) to the right-hand side yields

$$\frac{\partial \eta}{\partial t} = F_\eta - \nabla \cdot (\eta \nabla \chi) + \mathcal{J}(\eta, \psi), \tag{3.52}$$

$$\frac{\partial \delta}{\partial t} + g \nabla^2 h = F_\delta + \nabla \cdot (\eta \nabla \psi) + \mathcal{J}(\eta, \chi) - \nabla^2 K, \tag{3.53}$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (\bar{h} \nabla \chi) = F_h - \nabla \cdot (h' \nabla \chi) + \mathcal{J}(h, \psi). \tag{3.54}$$

The corresponding implicit system is

$$\zeta = Z, \tag{3.55}$$

$$\delta + \tau g \nabla^2 h = D, \tag{3.56}$$

$$h + \tau \nabla \cdot (\bar{h} \nabla \chi) = H. \tag{3.57}$$

This system is identical to the corresponding system (3.23)–(3.25) for the vorticity/divergence form, except that only the divergent part of the velocity is being treated implicitly; the procedures for solving these two systems are essentially the same.

3.3.2 Coriolis terms treated implicitly

Similarly, if we treat the Coriolis terms implicitly, then the split form would then be

$$\frac{\partial \eta}{\partial t} + \nabla \cdot (f \nabla \chi) \qquad = F_\eta - \nabla \cdot (\zeta \nabla \chi) + \mathcal{J}(\eta, \psi), \tag{3.58}$$

$$\frac{\partial \delta}{\partial t} - \mathcal{J}(f, \chi) + g \nabla^2 h = F_\delta + \mathcal{J}(\zeta, \chi) + \nabla \cdot (\eta \nabla \psi) - \nabla^2 K. \tag{3.59}$$

$$\frac{\partial h}{\partial t} + \nabla \cdot (\bar{h} \nabla \chi) \qquad = F_h - \nabla \cdot (h' \nabla \chi) + \mathcal{J}(h, \psi). \tag{3.60}$$

The corresponding implicit system is

$$\zeta + \tau \nabla \cdot (f \nabla \chi) \qquad = Z, \tag{3.61}$$

$$\delta - \tau \mathcal{J}(f, \chi) + \tau g \nabla^2 h = D, \tag{3.62}$$

$$h + \tau \nabla \cdot (\bar{h} \nabla \chi) \qquad = H. \tag{3.63}$$

This system is identical to the corresponding system (3.35)–(3.37) for the vorticity/divergence form, except that only the divergent part of the velocity is being treated implicitly. The solution in the case where f and $\bar{h}$ are both constant is easier than before (since ζ appears only in the first equation) but otherwise the same. In the general case (f or $\bar{h}$ not constant) it should be possible to solve using a "forced velocity" as before.

3.4 Potential vorticity/divergence form

By combining the vorticity equation (3.17) with the continuity equation (1.2) we can obtain the potential vorticity equation

$$\frac{\partial q}{\partial t} + \nabla \cdot (q \mathbf{v}) = F_q := \frac{h F_\eta - \eta F_h}{h^2} \tag{3.64}$$

where $q := \eta / h$ is the potential vorticity. This can be coupled with the divergence equation (3.18) and the continuity equation (1.2) to form a complete system which can be discretized by a semi-implicit scheme in various ways. Here, we consider only the case where $\bar{h}$ is constant and the Coriolis terms are treated explicitly, in which the split form (with explicit terms on the right) is

$$\frac{\partial q}{\partial t} \qquad = F_q - \nabla \cdot (q \mathbf{v}), \tag{3.65}$$

$$\frac{\partial \delta}{\partial t} + g \nabla^2 h = F_\delta + \mathbf{k} \cdot \nabla \times (\eta \mathbf{v}) - \nabla^2 K, \tag{3.66}$$

$$\frac{\partial h}{\partial t} + \bar{h} \delta \quad = F_\delta - \nabla \cdot (h' \mathbf{v}). \tag{3.67}$$

The corresponding implicit problem takes the form

$$q = Q, \tag{3.68}$$

$$\delta + \tau g \nabla^2 h = D, \tag{3.69}$$

$$h + \tau \bar{h} \delta = H, \tag{3.70}$$

where q, δ, and h are values at the new time level and and Q, D, and H are computed explicitly from values at the old time level. Clearly (3.68) gives q explicitly, and (3.69) and (3.70) can be solved for δ and h exactly as done for (3.24) and (3.26).

4 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 $\Delta t \to 0$? Here we follow the analysis of [2] and [9].

4.1 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 (1.3) can be written in the form

$$\frac{1}{\Delta t}\sum_{j=0}^{m} c_j \psi^{n+1-j} = \sum_{j=0}^{m} a_j \mathcal{A}\left(\psi^{n+1-j}\right) + \sum_{j=0}^{m} b_j \mathcal{B}\left(\psi^{n+1-j}\right). \tag{4.1}$$

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 (4.1) 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 (1.4), for which

$$m = 1, \qquad \mathbf{c} = (1, -1), \qquad \mathbf{a} = (1, 0), \qquad \mathbf{b} = (0, 1), \tag{4.2}$$

the trapezoidal/leapfrog scheme (1.5), for which

$$m = 2, \qquad \mathbf{c} = \left(\tfrac{1}{2}, 0, -\tfrac{1}{2}\right), \qquad \mathbf{a} = \left(\tfrac{1}{2}, 0, \tfrac{1}{2}\right), \qquad \mathbf{b} = (0, 1, 0). \tag{4.3}$$

Each time step of the scheme (4.1) requires solving the implicit problem

$$\psi^{n+1} - \tau \mathcal{A}\left(\psi^{n+1}\right) = \mathcal{C}^n := \frac{1}{c_0}\sum_{j=1}^{m}\left[-c_j \psi^{n+1-j} + a_j \Delta t \mathcal{A}\left(\psi^{n+1-j}\right) + b_j \Delta t \mathcal{B}\left(\psi^{n+1-j}\right)\right], \tag{4.4}$$

which is precisely the form of (1.6) with $\tau = a_0/c_0$. 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. Software for implementing schemes of this form is available from the author.

4.2 Consistency, order, and truncation error

To be of any value, the scheme (4.1) must approximate the equation (1.3). This is measured by the *truncation error* $\tau(t;\Delta t)$, defined as the amount by which the true solution ψ of (1.3) fails to satisfy the discrete equation (4.1).[3] Assuming that $\psi \in C^{p+1}$ and $\mathcal{A}, \mathcal{B} \in C^p$ we can use Taylor expansions to show

$$\begin{aligned}\tau(t;\Delta t) :=& \frac{1}{\Delta t}\sum_{j=0}^{m} c_j \psi(t-j\Delta t) - \sum_{j=0}^{m} a_j \mathcal{A}\left(\psi(t-j\Delta t)\right) - \sum_{j=0}^{m} b_j \mathcal{B}\left(\psi(t-j\Delta t)\right) \\ =& \frac{1}{\Delta t}\left(\sum_{j=0}^{m} c_j\right)\psi(t) + \sum_{k=0}^{p}\left[\frac{1}{(k+1)!}\left(\sum_{j=0}^{m}(-j)^{k+1}c_j\right)\psi^{(k+1)}(t)\right. \\ & \left. - \frac{1}{k!}\left(\sum_{j=0}^{m}(-j)^k a_j\right)\mathcal{A}^{(k)}(t) - \frac{1}{k!}\left(\sum_{j=0}^{m}(-j)^k b_j\right)\mathcal{B}^{(k)}(t)\right](\Delta t)^k + o\left((\Delta t)^p\right)\end{aligned} \tag{4.5}$$

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 (4.5) this requires at least

$$\sum_{j=0}^{m} c_j = 0. \tag{4.6}$$

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 (1.3) in (4.5) shows that this requires

$$\frac{1}{(k+1)!}\sum_{j=0}^{m}(-j)^{k+1}c_j = \frac{1}{k!}\sum_{j=0}^{m}(-j)^k a_j = \frac{1}{k!}\sum_{j=0}^{m}(-j)^k b_j \tag{4.7}$$

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 (4.7) with $k = 0$, i.e.,

$$-\sum_{j=0}^{m} j c_j = \sum_{j=0}^{m} a_j = \sum_{j=0}^{m} b_j \tag{4.8}$$

must hold in addition to (4.6). If the method is of order p then the truncation error reduces to

$$\begin{aligned}\tau(t;\Delta t) =& \sum_{j=0}^{m}\left[\frac{(-j)^{p+1}}{(p+1)!}c_j - \frac{(-j)^p}{p!}a_j\right]\mathcal{A}^{(p)}(t)(\Delta t)^p \\ &+ \sum_{j=0}^{m}\left[\frac{(-j)^{p+1}}{(p+1)!}c_j - \frac{(-j)^p}{p!}b_j\right]\mathcal{B}^{(p)}(t)(\Delta t)^p + o\left((\Delta t)^p\right).\end{aligned} \tag{4.9}$$

[3] This τ has no connection with the constant τ in (1.6) and (4.4); hopefully no confusion should result.

From (4.9) we draw the following conclusions:

- If the implicit and explicit methods are consistent separately, then the resulting semi-implicit method (4.1) 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 $\mathcal{B}(\psi)$.

This last point is significant: it implies that when the terms treated implicitly contain little energy (as they should), then the overall accuracy of the semi-implicit scheme may be determined by the order of the method used for the explicit terms (the method used for the implicit terms may be lower-order).

4.3 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 (1.3) is well-posed, then for any fixed time t the solution $\psi(t)$ depends continuously on the initial data $\psi(0)$. The scheme (4.1) 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., [15]) that the method is zero-stable if and only if the roots $\{r_j\}_{j=1}^m$ of the *characteristic polynomial*

$$P_0(z) := \sum_{j=0}^{m} c_j z^{m-j} \tag{4.10}$$

satisfy the *root condition*

$$|r_j| \le 1 \quad \text{and} \quad |r_j| = 1 \text{ only if } r_j \text{ is a simple root.} \tag{4.11}$$

Note that zero-stability depends only on the constants c_j. Also, for a consistent method $P_0(1) = 0$ by (4.6), 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 (4.1) 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., [15]), this theorem also holds for the case of CLM methods [9]. 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.

5 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). Indeed, it makes no sense to even look for absolute stability unless the solutions of the continuous problem remain bounded as $t \to \infty$. From this point on, the term "stable" will refer to *absolute stability*.

While the question of convergence of semi-implicit schemes has been answered in considerable generality (see section 4), there is less known about their absolute stability. Ascher et al. [2] analyzed the stability of selected CLM schemes for convection-diffusion problems (where the diffusion is treated implicitly), and Kang [9] likewise treated semi-implicit Runge-Kutta (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 [9] 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 [5]. On the other hand, some CLM methods such as (1.5) 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.

5.1 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 $\psi(t) = \mathbf{u}(t) \in \mathbb{R}^N$ for some N. In the linear constant-coefficient case where $\mathcal{A}(\psi(t)) = A\mathbf{u}$ for a constant matrix $A \in \mathbb{R}^{N\times N}$, the equation to be solved is

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u}. \tag{5.1}$$

If A is diagonalizable, i.e., there exists a nonsingular matrix P such that $\Lambda = P^{-1}AP$ is diagonal, then the transformation $\mathbf{w} = P^{-1}\mathbf{u}$ transforms (5.1) into

$$\frac{d\mathbf{w}}{dt} = \Lambda \mathbf{w}, \tag{5.2}$$

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

$$\frac{d\psi}{dt} = \lambda \psi \qquad (\lambda \in \mathbb{C}). \tag{5.3}$$

This represents a single component of (5.2), 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 (5.1) and the test equation (5.3) 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 (5.1) is $\mathbf{u}(t) = Pe^{\Lambda t}P^{-1}\mathbf{u}(0)$, so

$$\|\mathbf{u}(t)\| \le \|P\| \left\|e^{\Lambda t}\right\| \left\|P^{-1}\right\| \|\mathbf{u}(0)\| \tag{5.4}$$

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

$$\|\mathbf{u}(t)\| \le \text{cond}(P) \, \|\mathbf{u}(0)\| \tag{5.5}$$

where $\text{cond}(P) := \|P\| \, \|P^{-1}\|$ 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 $\text{cond}(P) = 1$ in the l_2 norm, so the solution is not only bounded but satisfies $\|\mathbf{u}(t)\| \le \|\mathbf{u}(0)\|$, 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 $\text{cond}(P)$, 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 (5.3) 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., [8]). 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 (5.3). Thus, consider the linear constant-coefficient case where $\mathcal{A}(\psi(t)) = A\mathbf{u}$ and $\mathcal{B}(\psi(t)) = B\mathbf{u}$ for constant matrices $A, B \in \mathbb{R}^{N \times N}$. The equation (1.3) to be solved is then

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u} + B\mathbf{u}. \tag{5.6}$$

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 $\mathbf{w} = P^{-1}\mathbf{u}$ transforms (5.6) into

$$\frac{d\mathbf{w}}{dt} = \Lambda_A \mathbf{w} + \Lambda_B \mathbf{w}, \tag{5.7}$$

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

$$\frac{d\psi}{dt} = a\psi + b\psi \qquad (a, b \in \mathbb{C}). \tag{5.8}$$

This represents a single component of (5.2), 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., [11]) 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 2. 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.

5.2 Absolute stability regions

Any time differencing scheme applied to a particular problem for $\psi(t)$ such as (1.3) 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 (5.3), 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 (5.3) 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 when $\mathrm{Re}\,(\lambda\Delta t) \leq 0$, i.e., when the original problem has bounded solutions. The regions of absolute stability for the Forward (FOR), second-order Adams-Bashforth (AB2), Backward (BACK), and Trapezoidal (TRAP) schemes are shown in Fig. 2. 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*.

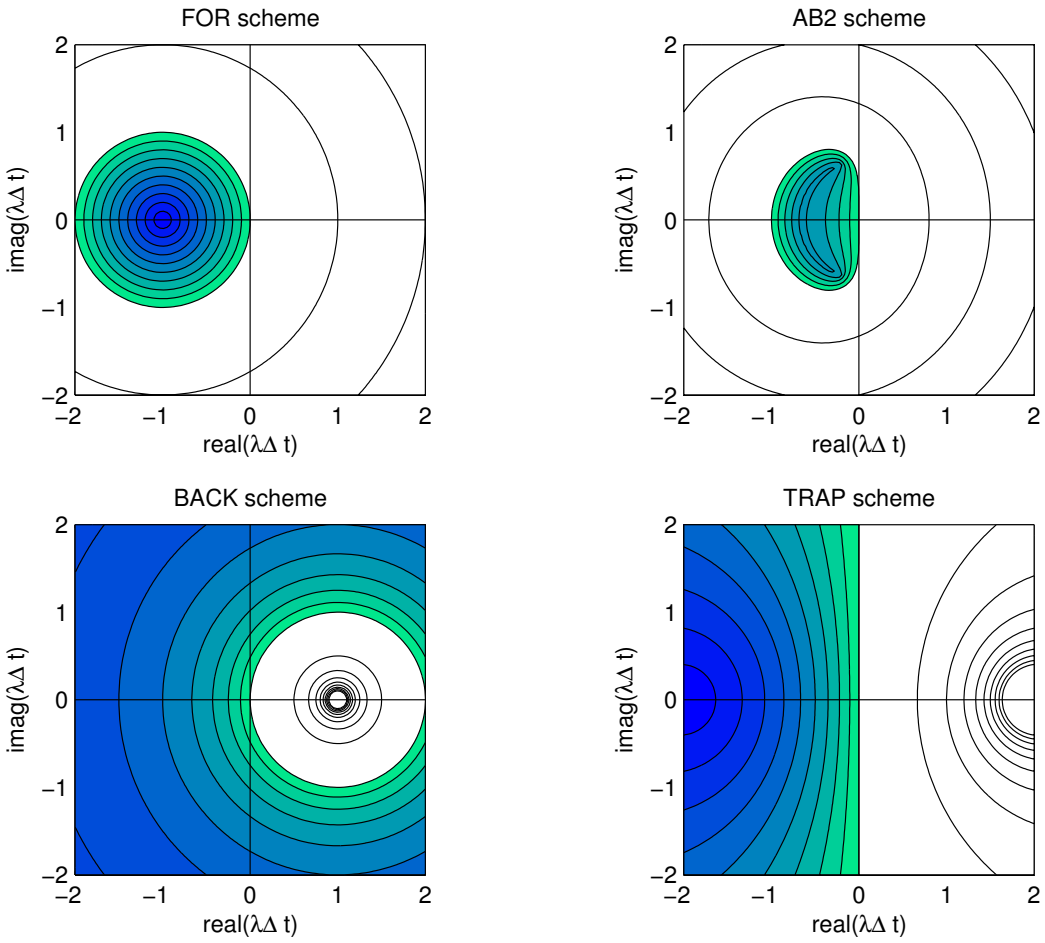

Figure 1: Absolute stability regions for four one-step schemes. The stability regions are shaded, and the contour interval for the amplification factors $r = \max_j |r_j|$ are 0.1 for $|r| \leq 1$ and 1.0 for $|r| > 1$.

To analyze a semi-implicit method we must use (5.8) as explained above. For this problem it turns out that the absolute stability depends on the products $\alpha = a\Delta t$ and $\beta = b\Delta t$. Specifically, any one-step semi-implicit method applied to (5.8) yields a solution of the form

$$\psi^n = [r(\alpha, \beta)]^n \psi^0 \tag{5.9}$$

where the complex number r is the amplification factor per time step.[4] For example, for the Backward/Forward scheme (1.4) we obtain

$$r(\alpha, \beta) = \frac{1+\beta}{1-\alpha}. \tag{5.10}$$

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

More generally, the CLM method (4.1) applied to (5.8) yields a solution of the form

$$\psi^n = \sum_{j=1}^{m} d_j [r_j(\alpha, \beta)]^n, \tag{5.11}$$

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

$$P(z; \alpha, \beta) := \sum_{j=0}^{m} (c_j - \alpha a_j - \beta b_j)\, z^{m-j} \tag{5.12}$$

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 (1.5) we have

$$r_{1,2}(\alpha, \beta) = \frac{\beta \pm \sqrt{1 + \beta^2 - \alpha^2}}{1 - \alpha}. \tag{5.13}$$

For each $j = 1, \ldots, m$, $r_j(\alpha, \beta)$ is the amplification factor for the corresponding solution mode in (5.11); the solution is bounded (for all initial conditions) if and only if these roots satisfy the root condition (4.11). The stability polynomial (4.10) and characteristic polynomial (5.12) 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 $r_j(0, 0) = r_j$, $j = 1, \ldots, m$). Thus, absolute stability implies zero stability but not conversely.[5] 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 $r_j(\alpha, \beta)$, $j = 2, \ldots, m$, 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 $j = 1, \ldots, m$.[6] Since in general we may regard a and b in the test equation (5.8) 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.

[5] 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.

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

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.

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.

5.3 One-step semi-implicit schemes

For a one-step scheme ($m = 1$), the requirements (4.6) and (4.8) for consistency reduce the general CLM method (4.1) to

$$\frac{\psi^{n+1} - \psi^n}{\Delta t} = \theta \mathcal{A}(\psi^{n+1}) + (1-\theta)\mathcal{A}(\psi^n) + \mathcal{B}(\psi^n), \tag{5.14}$$

where θ is the single free parameter. From (4.9) the corresponding truncation error is

$$\tau = \left[\frac{1}{2}\mathcal{B}'(t_n) + \left(\frac{1}{2} - \theta\right)\mathcal{A}'(t_n)\right]\Delta t + O\left((\Delta t)^2\right). \tag{5.15}$$

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

$$[\Omega_s + (1-\theta)\Omega_f]^2 \le \theta^2 \Omega_f^2. \tag{5.16}$$

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.

5.4 Two-step semi-implicit schemes

For a two-step scheme ($m = 2$), the requirements (4.6) and (4.7) for second-order accuracy reduce the general CLM method (4.1) to

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

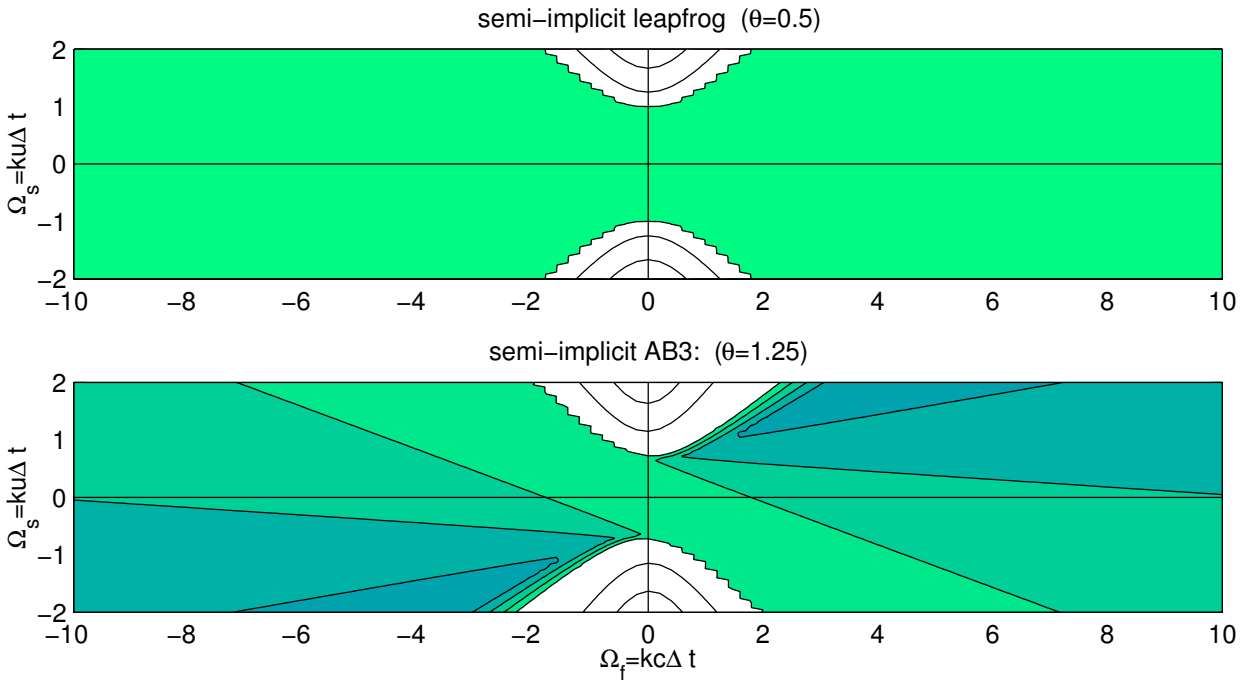

Figure 2: Amplification factors for the trapezoidal/leapfrog scheme (top panel) and the SI2/AB3 scheme (bottom panel) for oscillatory problems. The region of absolute stability is shaded, and the contour interval for the amplification factor is 0.1 for $|r| \le 1$ and and 1.0 for $|r| > 1$.

with two parameters c and γ left to specify. Some special cases include $(\gamma, c) = (0, 1)$ for the trapezoidal/leapfrog scheme (1.5) and $(\gamma, c) = (\frac{1}{2}, 0)$ for a trapezoidal/AB2 scheme.[7] An analysis shows that:

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

Thus, we conclude that of the family (5.17) of second-order two-step semi-implicit schemes, the traditional trapezoidal/leapfrog scheme (1.5) 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. 2.

[7] A modified version of this scheme—with an uncentered first-order implicit scheme—was used by [13].

5.5 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

$$\frac{\psi^{n+1} - \psi^n}{\Delta t} = a_0\mathcal{A}(\psi^{n+1}) + a_1\mathcal{A}(\psi^n) + a_2\mathcal{A}(\psi^{n-1}) + a_3\mathcal{A}(\psi^{n-2}) \\ + b_1\mathcal{B}(\psi^n) + b_2\mathcal{B}(\psi^{n-1}) + b_3\mathcal{B}(\psi^{n-2}). \tag{5.17}$$

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

$$b_1 = \tfrac{23}{12}, \qquad b_2 = -\tfrac{16}{12}, \qquad b_3 = \tfrac{5}{12}. \tag{5.18}$$

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)

$$a_0 = \theta, \qquad a_1 = \tfrac{3}{2} - 2\theta, \qquad a_2 = -\tfrac{1}{2} + \theta, \tag{5.19}$$

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 (5.17) with the coefficients (5.18) and (5.19) 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. 3. 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. 2.

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

$$a_0 = \theta, \qquad a_1 = \tfrac{23}{12} - 3\theta, \qquad a_2 = -\tfrac{16}{12} + 3\theta, \qquad a_3 = \tfrac{5}{12} - \theta, \tag{5.20}$$

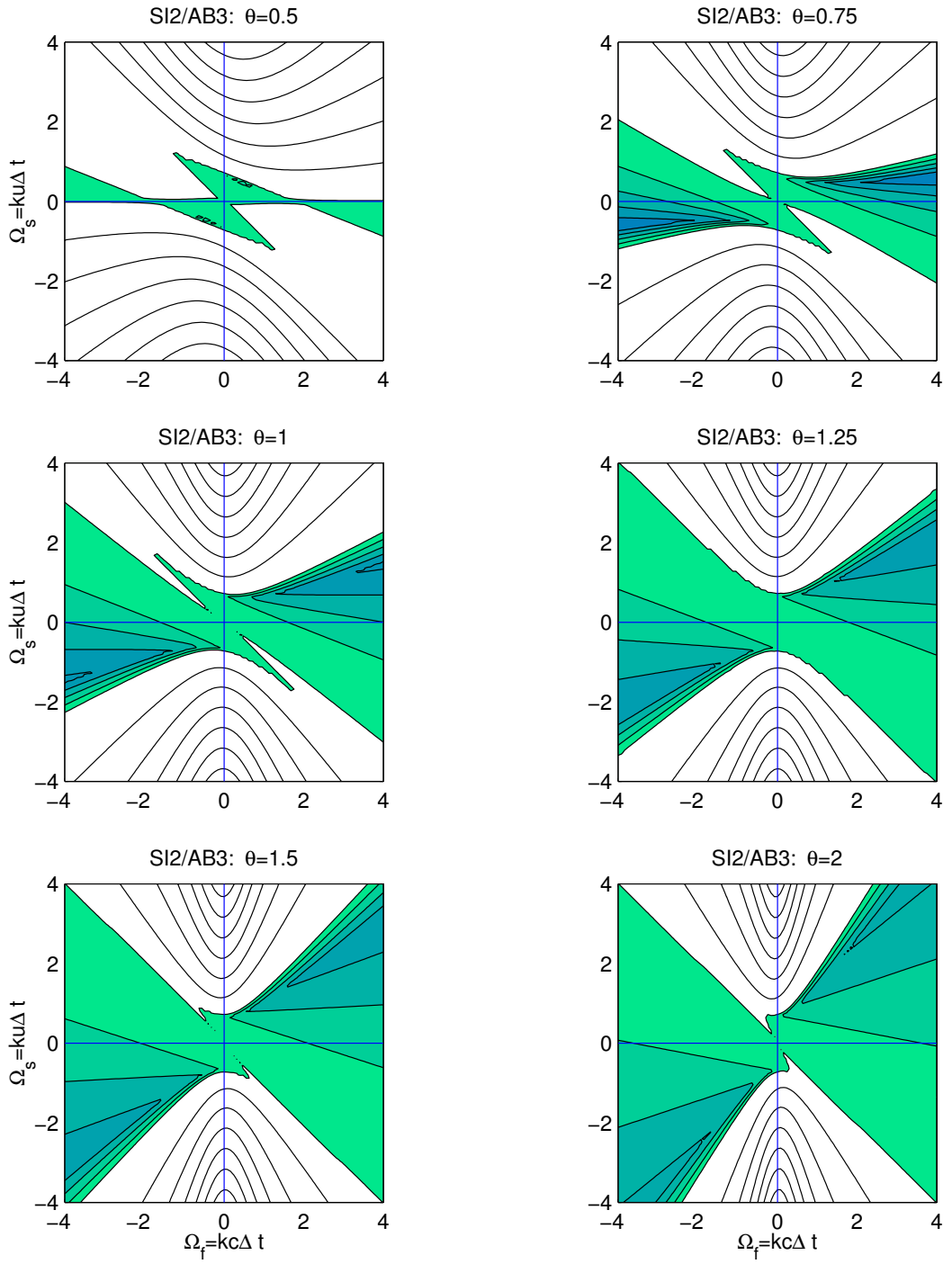

Figure 3: Amplification factors for the semi-implicit AB3 scheme for oscillatory problems. The six panels correspond to the values $\theta = 0.5$, 0.75, 1.0, 1.25, 1.5, and 2 as labeled. Contours and shading are as in Fig. 2.

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 (5.17) with the coefficients (5.18) and (5.20) 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.

6 Conclusions

We have reviewed the formulation semi-implicit time differencing schemes for the shallow-water equations. Analysis of the linearized equations identified the terms which should be treated implicitly. Splittings based on both the momentum and vorticity/divergence forms were developed. Even with a spatially varying reference depth and the Coriolis terms treated implicitly, in each formulation the implicit problem to be solved at each time step can be reduced to a linear elliptic equation (the vorticity/divergence form also requires solving Poisson problems for the velocity potential and streamfunction).

Analysis of general semi-implicit schemes based on linear multistep methods shows that combining explicit and implicit schemes which are both convergent leads to a convergent method. The overall order of the scheme is the minimum of the orders of the implicit and explicit methods used. However, if there is little energy in the modes treated implicitly then the overall accuracy may be that of the explicit method (which may be higher).

A search through part of the vast space of semi-implicit schemes 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 trapezoidal/AB2 scheme used (in modified form) by Simmons and Temperton [13] is unstable.
- The combination of the backward scheme with the third-order Adams-Bashforth (AB3) scheme is unstable.
- The combination of the trapezoidal scheme with the third-order Adams-Bashforth (AB3) scheme is unstable, as concluded by Durran [4]
- 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 the shallow-water model, and indeed for similar atmospheric models which admit fast and slow modes which are both oscillatory.

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] 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.

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

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

[6] Fulton, S. R., and W. H. Schubert, 1987: Chebyshev spectral methods for limited-area models. Part II: Shallow water model. *Mon. Wea. Rev.*, **115**, 1954–1965.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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