Reprinted from JOURNAL OF COMPUTATIONAL PHYSICS
All Rights Reserved by Academic Press, New York and London

Vol. 18, No. 4, August 1975
Printed in Belgium

Optimized Block-Implicit Relaxation

D. DIETRICH

Science Applications, Inc., McLean, Virginia 22101

AND

B. E. MCDONALD AND A. WARN-VARNAS

Naval Research Laboratory, Washington, D.C. 20375

Received November 5, 1974; revised February 19, 1975

A new relaxation method, block-implicit relaxation (BIR), which is applicable to partial difference equations with mesh varying coefficients and irregular boundaries, is compared with the less general Wachspress-optimized ADI method in solving the Poisson–Dirichlet problem on a rectangle. BIR consists of dividing a large computational mesh into several small meshes, and solving the difference equation exactly in each submesh interior. Residuals on the submesh boundaries are reduced by an iterative relaxation scheme. BIR is found superior for all but the largest forcing function scales. The large-scale convergence is accelerated significantly by a least-squares optimization procedure, which requires little additional computation or storage. In application to related sequences of problems for which accurate high-order extrapolation is possible, the new method has the strong advantage of performing such extrapolation with relatively little auxiliary storage or computation. Thus, the new method is well suited for time-implicit time marching models. Application to a time-implicit nonlinear transport equation with diffusion (high Reynolds' number channel flow) is discussed.

1. INTRODUCTION

In this paper, an efficient and general relaxation method for two-dimensional boundary value problems is described. The method can be succinctly described as "block-implicit relaxation (BIR)" in the same sense that alternating direction implicit (ADI) methods are described as "line-implicit relaxation." Actually, ADI methods are special cases of BIR, in which the implicit blocks are very elongated, spanning the entire mesh with only a few adjacent lines; in the case of second-order equations, three adjacent lines are used, with the outer two serving as temporarily fixed artifical "boundary conditions" for the interior line. The BIR method is most efficient for two-dimensional blocks, since a highly efficient

direct method is available for solving general two-dimensional problems. The latter method is a two-dimensional generalization of "shooting" methods for solving two-point boundary value problems and is termed the "generalized sweepout method (GSM)" by Hirota, Tokioka, and Nishiguchi [4]. However, as noted by Roache [11] (who discovered the GSM at about the same time) and by McAvaney and Leslie [6], the GSM cannot readily be applied as a direct method in solving large (high-resolution) problems, since this requires very high-precision arithmetic. Only by dividing such large problems into smaller ones, as accomplished with BIR, is the GSM readily applied to large problems. (Edwards and Hansen [3] describe a method for solving eigenvalue problems, which is related to the GSM in that a similar recursion relation is used. They use "conditioning transformations" to control round-off error. Such transformations are useful for homogeneous problems. However, for inhomogeneous problems considered in this paper, the necessary source term transform requires excessive computation to calculate and use. Therefore, it appears to us more advantageous to use the iterative procedure described in this paper in solving inhomogeneous problems.) On the other hand, BIR's efficiency is improved by using the GSM, especially in solving two-dimensional problems. Thus, *BIR and the GSM complement one another, and, as supported by the examples and discussion in this paper, make an attractive combination.*

The BIR–GSM combination is applicable to solving linear two-dimensional boundary value problems involving coupled partial difference equations with mesh varying coefficients and irregular[1] *boundaries. Perhaps the strongest points of the BIR–GSM procedure are its general applicability and BIR's highly efficient extrapolation capability.* Because of these points, and BIR's especially rapid convergence to small-scale solution components, the BIR–GSM procedure appears well suited for application to time-implicit formulations of time marching field problems, such as the example in Section 4.

2. The BIR Method

BIR consists of dividing a large computational mesh into several small meshes, and solving the difference equations exactly in each submesh interior. Residuals on the submesh boundaries are reduced by an iterative relaxation procedure. Additional BIR details are given by Dietrich [2] and by the examples below.

The GSM is recommended as a general method for solving two-dimensional

[1] In problems with irregular boundaries (not coincident with coordinate lines), at least some of the BIR blocks must be irregular. As noted by Roache [11], such blocks can be solved directly with the GSM.

BIR submesh problems; it is especially computationally efficient for sequences of problems with the same linear operators and class of boundary constraints. For such sequences, the BIR preprocessing is performed once and for all and requires $O(n^3)$ operations for each nXn submesh used. (If the linear operators and class of boundary constraints do not vary from one submesh to the next, the preprocessing need be performed on only one submesh.) After preprocessing, the GSM requires $O(n^2)$ operations for each $n \times n$ submesh solved, while other direct methods require $O(n^2 \ln n)$ operations. Iterative methods, such as ADI with a tridiagonal algorithm, require $O(n^2)$ operations per iteration. Thus, the GSM has clear advantages over other direct and iterative methods, and is the preferred method for many BIR applications. For a description of the GSM, see the Appendix.

In certain special problems, efficient direct methods are applicable even with high resolution. However, the BIR–GSM iterative approach is more general and may even be preferable, especially if high-precision results are not required. Even when high precision is required, the highly efficient BIR–GSM iteration may be preferred over an $O(n^2 \ln n)$ direct method.

Finally, the most efficient ADI methods are less general than the BIR–GSM method and do not have BIR's highly efficient extrapolation capability (see below) when applied to the total grid. Also, as noted in Section 6, ADI methods can be singular in problems where the GSM is not.

Returning to the BIR discussion, we assume (except as noted) that the BIR submesh boundaries are not altered from one iteration to the next and that the maximum submesh size compatable with the available computer precision is used (as recommended by Dietrich [2]). In such a case, the submeshes must overlap in order to be coupled during the BIR iteration. For second-order equations, we suggest a two-line overlap (Fig. 1). Although larger overlaps could be used, this would increase the calculation per BIR iteration (due to multiple updating of points in overlap regions), and would increase the auxiliary calculation and storage required in taking advantage of the following important BIR properties.

(i) The computation during one sweep depends only on values initially assumed on the mesh of block boundary lines (Fig. 1), so the exact solution can be calculated *in one sweep* from knowing only its exact values at a *small* subset of the total grid.

(ii) At the end of any sweep, the residuals are nonzero only on a mesh of lines adjacent to the block boundary lines (Fig. 1), which is again a *small* subset of the total grid.

The first feature allows one to extrapolate previous results while using only the values along the boundary lines. Thus, in a related sequence of problems, one can use high-order extrapolation with little auxiliary storage or computation. Together,

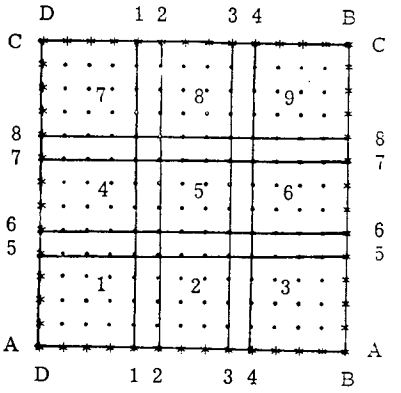

FIG. 1. Overlapping block arrangement for solving Poisson equation $\nabla_{ij}^2\phi \equiv (\phi_{i+1,j} + \phi_{i,j+1} + \phi_{i-1,j} + \phi_{i,j-1} - 4\phi_{i,j})/\Delta^2 = q_{ij}$, $2 \leqslant i \leqslant 13$, $2 \leqslant j \leqslant 13$ on a 14×14 grid, using nine 6×6 blocks. The boundary conditions are applied on the outer lines, *A*, *B*, *C*, and *D*. Sweeping left to right, bottom to top, as indicated by the block numbers (in the center of each block), the solution at the end of a given cycle depends only on the ϕ values assumed on the *even*-numbered lines at the beginning of the cycle. At the end of any cycle, the residuals ($\nabla_{ij}^2\phi - q_{ij}$) vanish everywhere except on the *odd*-numbered lines.

the two features allow one to combine previous trial solutions (generated by relaxation sweeps) in such a way that the mean squared residual is minimized, again with very little auxillary storage or computation; this allows one to optimize the convergence rate. Thus, both features allow one to optimize starting values and convergence rate with little auxillary storage or computation.

We now describe in detail the solution procedure for solving the discrete Poisson equation on the 14×14 grid illustrated in Fig. 1. using nine 6×6 blocks.

First, ϕ values are assigned along the even-numbered boundary lines (excluding the endpoints, where the true boundary constraints are applied). The accuracy at the end of the first sweep depends only on these values, so it is desirable to extrapolate them accurately from previous results if available. Next, the 16 interior ϕ values in block 1 (the lower left 6×6 block, bounded by lines A, D, 2, and 6) are adjusted so that the difference equation is satisfied *exactly* at these same points. The 16 ϕ values can be determined very efficiently using the GSM. At this point,

$$r_{ij} \equiv \nabla_{ij}^2\phi - q_{ij} = 0, \qquad 2 \leqslant i \leqslant 5, \quad 2 \leqslant j \leqslant 5 \tag{1}$$

Next, the same procedure is applied to block 2 (bounded by lines A, 1, 4, and 6), whose left boundary has four of the newly determined values from the first block. Note that, in adjusting the left four interior points (on line 2) of the second block, nonzero residuals are created at the right four interior points of the first block (on line 1). At this point,

$$r_{ij} = 0, \qquad 2 \leqslant i \leqslant 9 \ (i \neq 5), \quad 2 \leqslant j \leqslant 5.$$

The sweep is continued to the remaining seven blocks in the sequence indicated by the numbering, thereby completing the first grid sweep. At this point,

$$r_{ij} = 0, \qquad 2 \leqslant i \leqslant 13 \ (i \neq 5, 9), \quad 2 \leqslant j \leqslant 13 \ (j \neq 5, 9). \tag{2}$$

The residuals at the interior points on the four *odd*-numbered lines (Fig. 1), corresponding to r_{5j} , r_{9j} , r_{i5} , r_{i9} are then calculated and stored. There is no need to store the remaining residuals, which all vanish according to Eq. (2). If the residuals are too large, the ϕ values along the *even*-numbered lines are stored for future reference and a second sweep is performed. Again, the residuals are stored and, if too large, a third sweep may be started. However, it is desirable to use the available information to improve the ϕ values on the *even*-numbered lines before performing a third sweep. This is done by determining and using the linear combination of the first two sweeps that minimizes the mean squared residual along the *odd*-numbered lines, subject to the constraint that the residuals at all other interior points remain zero (see Eqs. (3)–(5) below). The stored ϕ values from the end of the first sweep are accordingly combined with the present values on the *even*-numbered lines to obtain the "optimum" starting values for the third sweep. However, the present ϕ values on the *even*-numbered lines are stored before performing the third sweep. When the third sweep is completed, three sets of residuals are available from which to optimize, if a fourth sweep is desired.

The computational details of the least-squares optimization referred to above are as follows. Let

$$\Phi_{ij} = \sum_{n=1}^{N-1} c_n \phi_{ij}^n + \left(1 - \sum_{n=1}^{N-1} c_n\right) \phi_{ij}^N, \tag{3}$$

where N is the total number of sweeps completed and the ϕ superscripts denote the sweep number. By construction, $r_{ij}^n \equiv \nabla_{ij}^2 \phi^n - q_{ij}$ vanishes everywhere except on the *odd*-numbered lines in Fig. 1; it follows that $R_{ij} \equiv \nabla_{ij}^2 \Phi_{ij} - q_{ij}$ will, according to Eq. (3), also vanish except on the odd lines for any choice of the c_n's. To determine the ideal c_n's, we render the mean squared residual associated with Φ stationary with respect to each c_n . That is, $(\partial/\partial c_n) \sum_{i,j} R_{ij}^2 = 0$, $n = 1, N - 1$, or

$$\frac{\partial}{\partial c_n} \left\{ \sum_{i,j} \left[\sum_{n=1}^{N-1} c_n \nabla_{ij}^2 \phi^n + \left(1 - \sum_{n=1}^{N-1} c_n\right) \nabla_{ij}^2 \phi^N - q_{ij} \right]^2 \right\} = 0, \qquad n = 1, N - 1. \tag{4}$$

Again, note that we need consider the i, j summation only for points on the *odd*-numbered lines. To rewrite Eq. (4) in terms of the stored residual vectors. we replace $\nabla_{ij}^2 \phi^n$ by $r_{ij}^n + q_{ij}$ and carry out the differentiation to get

$$\sum_{m=1}^{N-1} c_m \sum_{i,j} [(r_{ij}^N)^2 + r_{ij}^m r_{ij}^n - r_{ij}^m r_{ij}^N - r_{ij}^n r_{ij}^N] = \sum_{i,j} [(r_{ij}^N)^2 - r_{ij}^n r_{ij}^N], \quad n = 1, N - 1. \tag{5}$$

Thus, the optimum c_m's depend only on the self- and cross-correlations of the previously calculated and stored residual vectors. Solving Eqs. (5) for the c_m's and substituting into Eq. (3) for points on the *even*-numbered lines thus determines the desired optimum starting values for sweep number $N + 1$. Using such optimum ϕ values accelerates the convergence in general, giving a very significant improvement for the largest scales (more than doubling their convergence rate).

3. BIR Comparison with Wachspress-Optimized ADI

The Wachspress-optimized ADI method (Wachspress [13]) and the least-squares optimized BIR method are applied to the following sequence of discrete Poisson–Dirichlet problems.

$$\nabla^2_{ij}\phi^{mn} \equiv (\phi^{mn}_{i+1,j} + \phi^{mn}_{i\,j+1} + \phi^{mn}_{i-1,j} + \phi^{mn}_{i\,j-1} - 4\phi^{mn}_{i\,j})\,\Delta^{-2} = q^{mn}_{i\,j}$$
$$(1 \leqslant m \leqslant 30,\ 1 \leqslant n \leqslant 30), \qquad (6)$$

where

$$q^{mn}_{i,j} = \sin\left(2\pi m\,\frac{i-1}{61}\right) \cdot \sin\left(2\pi n\,\frac{j-1}{61}\right)$$

and

$$\phi^{mn}_{1\,j} = \phi^{mn}_{62,j} = \phi^{mn}_{i\,1} = \phi^{mn}_{i\,62} = 0 \qquad (2 \leqslant i \leqslant 61,\ 2 \leqslant j \leqslant 61).$$

The Wachspress method is of the Peaceman–Rachford class of ADI methods (Peaceman and Rachford [9])

$$(w_n - \delta_{xx})\,\hat{\phi}^{n+1}_{i,j} = (w_n + \delta_{yy})\,\phi^n_{i,j} - q_{i,j},$$
$$(w_n - \delta_{yy})\,\phi^{n+1}_{i,j} = (w_n + \delta_{xx})\,\hat{\phi}^{n+1}_{i,j} - q_{i,j}, \qquad n = 1, 2, \ldots, 2^P,$$

where $2 \cdot 2^P$ line-implicit sweeps are performed, thereby completing 2^P iterations; w_n are the Wachspress-optimized iteration parameters; $\delta_{xx}\phi = (\phi_{i-1,j} + \phi_{i-1,j} - 2\phi_{i,j})\,\Delta^{-2}$; and $\delta_{yy}\phi = (\phi_{i,j+1} + \phi_{i,j-1} - 2\phi_{i,j})\,\Delta^{-2}$. The iteration parameters are determined by minimizing the maximum possible (over all eigenfunctions) error ratio (final error over initial error) that results after completing 2^P iterations. This error ratio is given by

$$\prod_{n=1}^{2^P} (\lambda_k - w_n)(\lambda_k + w_n)^{-1},$$

where λ_k is the kth eigenvalue. For problems in which all eigenvalues are real and of the same sign, such as the present one, it is always possible to determine iteration parameters which give error ratio less than unity for all eigenfunctions.

For convenient comparison of the Wachspress and BIR methods, we define a normalized mean squared residual, E^{mn}:

$$E^{mn} \equiv \sum_{i=2}^{61} \sum_{j=2}^{61} (\nabla_{ij}^2 \phi^{mn} - q_{i,j}^{mn})^2 \cdot \left[\sum_{i=2}^{61} \sum_{j=2}^{61} (q_{i,j}^{mn})^2 \right]^{-1}. \tag{7}$$

In Fig. 2, the two methods are compared in detail in solving Eq. (6) after roughly equal amounts of computation: about 50 operations per grid point. This corresponds to four iterations of the four-parameter Wachspress method and four sweeps of the BIR method. The plotted number in the (m, n) position of the Fig. 2 diagrams is the base ten logarithm of $(E^{mn})^{-1}$. Thus, large numbers reflect small residuals.

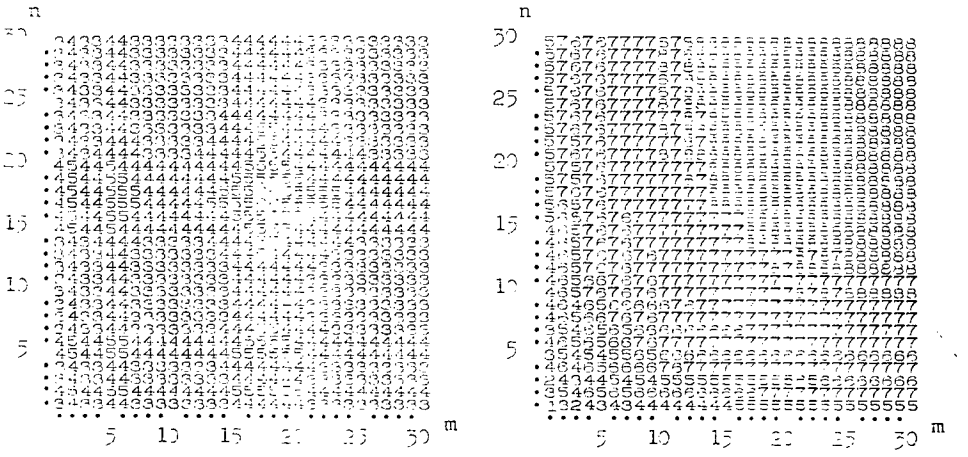

FIG. 2. Wavenumber space diagrams revealing Wachspress-optimized ADI and optimized BIR convergence rates for all possible two-dimensional wavenumber source terms, q^{mn}, in solving discretized Poisson–Dirichlet problem on a 62×62 grid, $\nabla_{ij}^2 \phi^{mn} \equiv (\phi_{i-1,j}^{mn} + \phi_{i,j+1}^{mn} + \phi_{i-1,j}^{mn} + \phi_{i,j-1}^{mn} - 4\phi_{i,j}^{mn})\Delta^{-2} = q_{ij}^{mn} \equiv \sin(2\pi m((i-1)/61)) \cdot \sin(2\pi n((j-1)/61))$. Defining the normalized mean squared residual residual, E^{mn}, as $E^{mn} \equiv \sum_{i=2}^{61} \sum_{j=2}^{61} (\nabla_{ij}^2 \phi^{mn} - q_{ij}^{mn})^2 \cdot [\sum_{i=2}^{61} \sum_{j=2}^{61} (q_{ij}^{mn})^2]^{-1}$, *the integer plotted in the (m, n) diagram position is the negative of the base ten logarithm of E^{mn}, rounded off to the nearest integer, evaluated after four relaxation sweeps.* Both methods require about 50 operations per grid point total to complete four sweeps. All wavenumber (m, n) problems are initialized with $((\phi_{ij}^{mn} = 0, i = 1, 62), j = 1, 62)$, corresponding to $E^{mn} = 1$ initially. Thus, large integers reflect rapid convergence rates.

The diagrams show that, with the nearly equal amounts of computation used, *the BIR method has smaller error than the Wachspress-optimized ADI method for all but the largest scales.* Not surprisingly, BIR is superior for scales smaller than the block size used ($m > 4$, $n > 4$). (In these results 16 17×17 blocks were used. Of course as noted by Dietrich [2], convergence is maximized by using the largest block size compatible with the available computer precision.)

Figure 2 contains the basic information regarding BIR performance, although

detailed phase information is not revealed. Only the phase relative to the internal block boundaries is important; the exact solution results after one sweep if the block boundaries are all nodes of the solution. Thus, BIR effectively becomes a direct method for such an idealized case. (The Wachspress method can also be made direct for an idealized case when the source term is exactly one of the eigenfunctions; this is done by using the iteration parameter appropriate for that eigenfunction. However, in contrast to the idealized BIR case, using such an iteration parameter would cause *large* errors in other source term components that may be present in a real problem.) Since the phase (relative to the block boundaries) varies nearly randomly in the two-dimensional wavenumber space of the results displayed in Fig. 2, *the BIR results are approximately representative.* Some wavenumbers would converge faster if the phase were changed arbitrarily, and some would converge more slowly. (This has been verified by creating the same type of diagram for cosine functions in Eq. (6).)

Caution should be taken to properly interpret the word "scale" in the above discussions. "Scale" refers strictly to the eigenfunction wavelength. If the source term were like a *narrow* Gaussian function, the half-width corresponds to the dominant scale only for the first few BIR sweeps, during which the rapid small-scale convergence indicated in Fig. 2 occurs. After the first few BIR sweeps, the error is very small; however, it is concentrated in the long wavelength eigenfunction components of the source term. *Thus, after the first few BIR sweeps, the convergence rate for the narrow Gaussian source term is reduced by the slower convergence of its long wavelength components, as revealed in Fig. 2, although the error is already very small.*

Although the Wachspress method is competitive with BIR in some respects for the present problem, the BIR method is more versatile and, being more constrained by numerical precision, would gain relative to the Wachspress method when higher-precision arithmetic is used (along with fewer blocks). Obviously, BIR is superior when sufficiently high-precision arithmetic is used, with one (or very few) block submesh(es).

The slower convergence of large scales with the BIR method should not be a serious drawback in solving the partial difference equations arising from time-implicit models in which small scales vary most rapidly in time. Large scales, being of primary interest in most problems, are usually well resolved in time and are stable in time even for explicit time schemes. Further, as they are well resolved in time, they can be extrapolated to high order and accuracy. *Thus, the slower BIR convergence for large scales can be compensated, at least partially, by its highly efficient extrapolation capability in time marching problems.* (Small scales, on the other hand, may change greatly during a model time step and cannot be well extrapolated. In fact, when the model CFL condition is violated, small scales lose time continuity, so it may be desirable to filter high frequencies before performing

any high-order extrapolation. As is well known, the small scales usually are numerically unstable in time under such conditions, unless they are converged in a time-implicit model.)

Finally, another possible refinement of BIR is to alter the block boundaries after each sweep. This has been found beneficial in solving the Poisson equation with doubly periodic boundary conditions but has not been combined with or compared with the least-squares optimization described above. It also appears possible that using higher-order difference approximations near interior boundary lines may accelerate the convergence rate.

4. An Example: BIR Application to a Time-Implicit Model of High Reynolds' Number Flow in a Channel

The BIR procedure, using the GSM, has been applied to forced two-dimensional incompressible high Reynolds' number flow in a channel. Using a fully time-implicit formulation of the barotropic vorticity equation governing the flow—in which the nonlinear terms are resolved into a linearized part plus a nonlinear part—the resulting time marching problem is numerically stable, even when the advective CFL condition is exceeded by a large factor.

Before describing the fully implicit procedure, we momentarily consider the conditions for which time-implicit formulation, which allows one to exceed CFL conditions, is advantageous.

In simulating time-dependent field phenomena which are "quasisteady" in the sense that local time tendencies are calculated from sums whose individual terms are large compared to their sum, the effect of any small relative error in representing a given term can be greatly magnified to produce severe tendency errors. Thus, to get a reliable forecast (from the tendencies), one must approximate each contributing term with high accuracy. Physical processes must be accurately represented in the model equations and any space finite difference derivative approximations which are used must be very accurate. Such accurate derivative approximations may be determined by the following methods.

(i) Using fine space resolution, with grid interval much smaller than the "quasisteady" space scales of interest;

(ii) Using high-order difference approximation;

(iii) Using the pseudospectral method (Merilees [7]).

The latter two methods can accomplish a given accuracy while using lower space resolution than required using the first method. However, they require more computation per time step per retained grid point.

The first method (fine space resolution) usually requires more storage, and more seriously, requires time resolution of high-frequency small-scale modes resolved by the grid unless a time-implicit formulation is used. If such modes are of small amplitude and the fine space resolution is desired only to obtain accurate, relatively large-scale derivatives, it is then desirable to use an implicit approach. In any event, certain small-scale modes may have exceptionally short time scales, and resolving their time dependence may be of little value to the problem at hand. (For example, in atmospheric forecast models, small-scale gravity waves propagating relative to a fast current have exceptionally high frequencies, and resolving their time dependence is of little value—except, perhaps, their more slowly varying large-scale transport properties which, hopefully, can be adequately parameterized. It should also be noted that certain fine structures such as fronts have much longer time scales, so exceeding the model CFL condition does not necessarily ruin their numerical forecasts.) Thus, the desirability of implicit time marching is well established. For further details, see Kwizak and Robert [5] and O'Brien and Hurlburt [8].

Implicit time marching usually requires at least approximate solution of partial differential equations each time step. These equations can be very complicated and generally have space varying coefficients. To put the fine space resolution method in better comparison with the other two, better methods for solving such equations are needed. In the present example, a general method for solving such equations is described and applied to a simple problem.

The procedure is as follows, starting with the barotropic vorticity equation and using a stream function ψ for the nondivergent velocity.

$$\frac{\partial \zeta}{\partial t} + \nabla \cdot \zeta \mathbf{V} - \nu \nabla^2 \zeta = 0, \qquad |x| < \frac{L}{2}, \qquad |y| < \frac{D}{2}, \tag{8}$$

where

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y}; \qquad \nabla^2 = \nabla \cdot \nabla; \qquad \zeta = \nabla^2 \psi; \qquad \mathbf{V} = u\mathbf{i} + v\mathbf{j};$$

$$u = -\frac{\partial \psi}{\partial y}; \qquad v = \frac{\partial \psi}{\partial x};$$

and ν is a specified diffusion coefficient. The boundary conditions are

$$\begin{aligned} &u(x, \pm D/2, t) = v(x, \pm D/2, t) = v(\pm L/2, y, t) = 0, \\ &u(\pm L/2, y, t) = U(y^2 - (D/2)^2)(1 - \epsilon y), \end{aligned} \tag{9}$$

where U is a specified flow rate and ϵ eliminates the possibility of the trivial

steady state $u = U(y^2 - (D/2)^2)$ for which all terms vanish. If ϵ is large, the possibility of strong barotropic energy exchange between the x-averaged flow and disturbance exists. The initial conditions are

$$\psi(x, y, 0) = \int_0^y U(y^2 - (D/2)^2)(1 - \epsilon y)\, dy. \tag{10}$$

This problem is used only to illustrate BIR–GSM application to time-implicit models and is chosen for its simplicity. Equation (8) is approximated by the time-implicit difference equation

$$\zeta^{t+\Delta t} - \zeta^t = \Delta t(-\nabla \cdot \bar{\zeta}\bar{\mathbf{V}} + \nu\nabla^2\bar{\zeta}), \tag{11}$$

where $\bar{\zeta} = \frac{1}{2}(\zeta^t + \zeta^{t+\Delta t})$ and $\bar{\mathbf{V}} = \frac{1}{2}(\mathbf{V}^t + \mathbf{V}^{t+\Delta t})$. Equation (11) is within $O(\Delta t)^2$ of the Crank–Nicolson (or trapezoidal) approximation, so it also has error $O(\Delta t)^2$. In the absence of space truncation error, this time-implicit formulation conserves enstrophy (squared vorticity) exactly when $\nu = 0$ and $\nabla \cdot \mathbf{V} = 0$ (Roberts [12]). The latter relation is satisfied by the use of the stream function; if it were not, exact enstrophy conservation could be retained by replacing $\nabla \cdot \bar{\zeta}\bar{\mathbf{V}}$ by $\frac{1}{2}[\nabla \cdot \bar{\zeta}\bar{\mathbf{V}} + (\bar{\mathbf{V}} \cdot \nabla)\bar{\zeta}]$ in Eq. (11) (Piacsek and Williams [10]).

Letting

$$\begin{aligned} \mathbf{V}(x, y, t) &= \mathbf{V}(x, y, t_0) + \mathbf{V}'(x, y, t) \equiv \mathbf{V}_0 + \mathbf{V}', \\ \zeta(x, y, t) &= \zeta(x, y, t_0) + \zeta'(x, y, t) \equiv \zeta_0 + \zeta', \end{aligned} \tag{12}$$

Eq. (11) may be rewritten as

$$[\zeta' - \mathscr{L}(\mathbf{V}_0, \mathbf{V}', \zeta_0, \zeta')]^{t+\Delta t} = [\zeta' + \mathscr{L}(\mathbf{V}_0, \mathbf{V}', \zeta_0, \zeta')]^t + Q_0 + Q', \tag{13}$$

where $Q_0 = \Delta t[-\nabla \cdot (\mathbf{V}_0\zeta_0) + \nu\nabla^2\zeta_0]$; $Q' = \Delta t[-\nabla \cdot (\bar{\mathbf{V}}'\bar{\zeta}')]$; and

$$\mathscr{L}(\mathbf{V}_0, \mathbf{V}', \zeta_0, \zeta') = (\Delta t/2)[-\nabla \cdot (\mathbf{V}_0\zeta' + \mathbf{V}'\zeta_0) + \nu\nabla^2\zeta'].$$

Finally, centered second-order space differencing is used. Although the resulting space differenced nonlinear transport terms do not converse enstrophy exactly, the finite difference analog of replacing $\nabla \cdot \bar{\zeta}\bar{\mathbf{V}}$ by $\frac{1}{2}[\nabla \cdot \bar{\zeta}\bar{V} + (\bar{\mathbf{V}} \cdot \nabla)\bar{\zeta}]$ would lead to *exact* (algebraic) enstrophy conservation by these terms in the space *and* time differenced equation. Analogous quadratic conservation is possible with time-implicit primitive equation models (Dietrich [15]).

For some time after $t = t_0$, Eq. (13) is dominated by the linear terms and Q_0, with perturbation product terms, Q', being relatively small. The coefficients of the linear terms vary in space but not in time (except when relinearization is performed).

The solution procedure is as follows. To initiate each time step, the square-bracketed term on the right-hand side of Eq. (13) is calculated using previous time

step results. Then, Q' is guessed or extrapolated from previous time step results. Then, the resulting linearized partial difference equation is solved for a first approximation of the future stream function, using BIR and the GSM. (The term Q_0 is calculated and stored only when it is adjusted during relinearization.) It is then preferable to obtain a better approximation for the initially guessed Q' term by substituting the first-approximation stream function. The process is iterated until sufficient convergence is attained. If the iterations start to diverge, or converge too slowly, this indicates that the Q' terms have grown too large for the time step being used. This is a signal to redefine the basic state variables ($\mathbf{V}_0$, ζ_0) to be the present flow, appropriately adjust the right-hand side of Eq. (13), and restart the predictor–corrector sequence. Very rapid convergence results. Finally, the next time step is initiated.

Although Eq. (13) may be viewed as a fourth-order equation for the stream function, the recursion relation for the GSM can be broken into two second-order ones, involving the vorticity and stream function. This requires storage of two rows of ζ and greatly simplifies the recursion. Both computational savings and program simplification result if the fields u and v are also stored.

5. Directional Reynolds' Number and Resolution

When $\epsilon = 0$, steady parabolic parallel flow is possible which is stable unless UD/ν is very large. Thus, for small ϵ, one would also expect nearly parallel flow to be maintained. It follows that the v-component flow, normal to the boundaries $y = \pm D/2$, also should remain small compared to U. In such a case, the y-directional Reynolds' number, $\mathrm{Re}_y = vD/\nu$, is relatively small everywhere, even though the channel Reynolds' number, $\mathrm{Re} = UD/\nu$, may be very large. This is desirable in GSM application to linearized transport equations, since the GSM recursion relation can be singular or nearly singular unless the local grid interval Reynolds' number for flow in the recursion direction is less than unity everywhere (see Appendix). In the present problem, such singularity is easily avoided by using the "slow flow" direction (the y-direction) as the GSM recursion direction and using sufficient y-resolution to guarantee satisfaction of the *grid interval Reynolds' number criterion*, $v\Delta y/\nu < 1$. (Alternatively, one could avoid this criterion by not solving the y-direction advection term implicitly.)

6. Channel Problem Results

To illustrate the numerical behavior of the procedure described in Section 4, results are discussed for three cases in which only the time step is varied. In cases 1, 2, and 3 the common parameters are: $U = 1$, $\nu = 0.01$, $L = 12$, $D = 2$, and

$\epsilon = 0.1$. The grid is 80×15 points and six (15×15) blocks are used. The boundaries lie between the outer two lines of grid points. The varied parameter, Δt, is, respectively, 2, 1, and 0.5. The channel Reynolds' number, UD/ν, is 200. In the first case, the horizontal advection CFL condition is exceeded by a factor of thirteen. The deviation from initial conditions is nearly the same function of time for all three cases; the agreement is within two percent at all discrete times of the first case, and agreement improves with time, as all three cases approach the same steady state. The misrepresented small scales do not spoil the long-term integration.

These results indicate that the time resolution is adequate for the dominant large-scale flow in all three cases. With increasing time step, the number of GSM sweeps required for the first perturbation product iteration increases somewhat. However, fewer than twice as many sweeps are needed for case 1 as for case 3 and, after the first iteration, one BIR sweep is adequate in all three cases. The same BIR convergence criterion (1 % average change on the interior block boundaries) is used in all three cases. This is very good behavior in spite of the unrefined procedure used: the block boundaries are fixed in space, least-squares optimization is not used, and the GSM sweeps are started with the most recently calculated values on the block boundaries rather than using higher-order extrapolation. Finally, the perturbation product iteration averages about two corrections per time step for all three cases, while using the same convergence criterion of 1 % average change and using linear extrapolation to start the iteration.

The applicability of the BIR–GSM to time-implicit atmospheric-type models has been further investigated by testing the convergence rate of a modified version of the time-implicit equation in Section 4. The basic differences are: the "side" boundary conditions have been replaced by periodic conditions; the horizontal eddy diffusion term is omitted; and the blocks are "staggered" in the periodic direction from one BIR sweep to the next. The results are very encouraging. All wavenumbers converge rapidly if the time step (only one step is performed) is not longer than the advective CFL based on the horizontal block dimension. For example, when the particle displacement is roughly half the block dimension in one time step, the normalized mean squared residuals decrease by a factor of roughly 300 per BIR sweep. For roughly one fourth block displacement, the factor is roughly 15,000.

Such a time-implicit problem with infinite Reynolds' number in one or more directions would not be well-suited for solution by ADI methods; when the linearized advection CFL condition is violated and the flow is not unidirectional in the infinite Reynolds' number direction, the associated linearized ADI matrix equation for that direction can be singular or nearly singular. To avoid such an undesirable situation when violating the CFL condition, one must either introduce dissipation to satisfy the directional grid interval Reynolds' number criterion, or transform the problem in a way that results in large time truncation error, unless

accurate high-order extrapolation of previous results is performed or the transformed ADI matrix equation is iterated. The latter requires extra computation and/or storage and may still be numerically unstable; the former may be undesirable for physical reasons. In an atmospheric model, satisfaction of the horizontal grid interval Reynolds' number criterion requires the use of unphysically large horizontal eddy viscosity and/or extremely high space resolution. If a horizontal grid interval of 300 km were used, the appropriate eddy viscosity for wind speeds of 30 m/sec would be at least 10^{11} cm^2/sec. (This would rapidly dissipate small-scale features, such as fronts, which might develop using more realistic eddy viscosity values. Further, rapid large-scale dissipation would occur due to *horizontal* eddy transfer parameterization, while most synoptic scale dissipation is probably due to *vertical* transfers in the surface boundary layer.) More realistically, horizontal eddy transfer should be based on the resolved horizontal wind *gradients* rather than magnitude; it should not attain large vales for broad but fast currents.

7. Further Remarks on BIR and Atmospheric Models

Perhaps one of BIR's most promising applications is to long-term global atmospheric circulation and climate models. In such models, it is desirable to greatly exceed the CFL condition, especially near the pole regions. Near spherical coordinate poles, the longitudinal resolution must be high—and, most naturally, is (due to the convergence of the meridions in polar regions)—in order to accurately simulate the polar regions (where the advection and curvature terms are amplified, due to the coordinate curvature, relative to local time tendencies). A "primitive equation" model can be devised which conserves energy and mass exactly (in the absence of physical sources or sinks). The highly desirable energy conservation property guarantees computational stability, even when exceeding the model CFL condition. Energy conservation requires using a fully implicit model whose coupled, nonlinear equations are probably most readily solved using a BIR–GSM procedure analogous to the one described in Section 4. In a three-dimensional model, one would probably apply the BIR–GSM combination to an "alternating direction plane-implicit" formulation, using only the x–z and y–z planes, and marching only vertically (in the z direction) with the GSM. This would retain the efficient time extrapolation capability of BIR and allow use of realistic horizontal and vertical eddy transfer coefficients, while exceeding the model CFL conditions. Such an implicit approach would put the use of low-order, relatively high-resolution grid models in better comparison with high-order and pseudospectral models.

However, due to their versatility, spectral or Galerkin methods may still be superior for simulating phenomena characterized by widely separated bands in wave number space which actively influence one another by direct nonlinear

interaction (as opposed to the nonlinear cascade mechanism). Powerful east coast winter storms, in which there is apparently strong direct interaction between the synoptic scale and the convective cloud scale, resulting from cold continental synoptic scale air masses rushing over warm water, might well be such a phenomenon. The basic physics of such band phenomena can be described in the framework of linearized instability theory.

Such phenomena would be well suited for simuation with a "spectral gap" model (Warn [14]) in which relatively inactive intermediate scales between the bands are ignored. Strong "interband" interaction would be most likely when the group velocity of a relatively small-scale band can approximate the phase velocity of another band. In such a case, the large scale could "modulate" the smaller scale while the nonlinear transport by the smaller scale could have a significant larger-scale effect. To eliminate numerically troublesome high frequencies usually associated with the smaller-scale phase speed, which are physically irrelevant to the larger scale, one could subtract out the mean frequency of the two smaller-scale components of each interacting triad. (For example, if $u = \cos\{kx + \omega_k t\}$ and $v = \cos\{[k + \Delta k]x + [\omega_k + \Delta k(\Delta\omega/\Delta k)]t\}$, $|\Delta k| \ll |k|$, the long wave component of $u \cdot v$ is given by

$$\{uv \mid \cos[\Delta kx]\} + \{uv \mid \sin[\Delta kx]\} = \tfrac{1}{2}\cos\{\Delta kx + \Delta k(\Delta\omega/\Delta k)t\}.$$

The same long wavelength result is obtained after replacing u by $\hat{u} = \cos\{kx - \frac{1}{2}\Delta\omega t\}$ and v by $\hat{v} = \cos\{[k + \Delta k]x - \frac{1}{2}\Delta\omega t\}$. Thus, one still obtains the correct nonlinear effect on the long wave component after greatly reducing the short wavelength frequencies, assuming $|\Delta\omega \ll \omega|$. Also, if the long wavelength component has phase speed close to $-\Delta\omega/\Delta k$, strong sustained interaction with the two short wave components can occur: if the product of the short wave components has a frequency close to the natural frequency of the long wave component, near-resonance results.) Finally, more detailed smaller-scale structure could easily be described by including harmonics of the smaller-scale band in the spectral model. Further details on the versatility of spectral and Galerkin models are discussed by Dietrich [1].

8. Conclusion

The optimized BIR method, combined with the GSM, appears well suited for two-dimensional boundary value problems, including higher-dimensional problems which have been reduced to sequential two-dimensional problems. Modes of scale smaller than the two-dimensional blocks used with BIR converge very rapidly, which is desirable in time marching problems for which small space scales also have

small time scales. Results from application to an implicit time formulation of the barotropic vorticity equation suggest that useful application to implicit atmospheric models is possible.

Finally, some relevant features of BIR and the GSM are as follows.

1. When combined with least-squares optimization (which requires little auxiliary calculation or storage, due to properties unique to BIR), convergence rate is superior to Wachspress-optimized ADI methods in solving the Poisson–Dirichlet test problem for all but the largest-scale forcing functions.

2. High-order extrapolation for starting relaxation sweeps requires little auxiliary storage or calculation. In problems for which accurate large-scale extrapolation of previous results is possible, this at least partially compensates for the slower BIR convergence rate to large-scale forcing function components.

3. When combined with the GSM, a single BIR sweep requires about the same amount of computation as an ADI sweep (using the tridiagonal algorithm) while, at most, doubling the storage; the BIR–GSM combination can be applied to a large class of problems, including linear coupled partial difference equations with variable coefficients (no dependent variable elimination is needed).

4. Although both BIR and the GSM generalize to an arbitrary number of dimensions, the GSM is highly efficient only for one- and two-dimensional problems.

5. The BIR–GSM combination appears especially well suited for problems with complicated geometry (at present, the most efficient ADI methods are restricted to simple geometries).

6. The GSM, besides being more general, can be competitive with other fast direct methods, even accounting for the higher-precision arithmetic required for large problems.

APPENDIX: THE GENERALIZED SWEEP-OUT METHOD (GSM)

The GSM generates the solution of finite difference boundary value problems by calculating and superposing two associated solution components: a particular solution which satisfies some of the boundary conditions and a homogeneous solution which cancels boundary condition errors usually occurring in the particular solution. Although, as noted by Roache [11] and McAvaney and Leslie [6], the GSM is basically unstable (i.e., highly sensitive to round-off error when applied *directly* to large problems), its high computational efficiency in application to small- or moderate-resolution problems makes its use with BIR both natural and, as supported by the examples and discussion in the text, of practical value.

As a direct method, the GSM is most efficient for solving two-dimensional problems. In particular, it applies to equations of the type

$$\nabla^2\phi + \mathbf{A}(x, y) \cdot \nabla\phi + B(x, y)\phi = S(x, y), \tag{A.1}$$

with any linear, properly posed boundary constraints specified on a closed boundary. GSM application is limited by numerical stability rather than by the functional forms of the specified coefficients $\mathbf{A}(x, y)$ and $B(x, y)$ in Eq. (A.1). In contrast, other efficient direct methods, such as those involving odd–even reduction or Fourier transforms, rely on $\mathbf{A}$ and B being of special form, such as being constant or having very few Fourier coefficients.

Equation (A.1) may be approximated as

$$\begin{aligned}\phi_{i,j+1}[\Delta y^{-2} + AY_{ij}/(2\Delta y)] + \phi_{i,j-1}[\Delta y^{-2} - AY_{ij}/(2\Delta y)] \\ + \phi_{ij}[-2(\Delta x^{-2} + \Delta y^{-2}) + B_{ij}] \\ + \phi_{i+1,j}[\Delta x^{-2} + AX_{ij}/(2\Delta x)] + \phi_{i-1,j}[\Delta x^{-2} - AX_{ij}/(2\Delta x)] \\ = S_{ij}\,; \qquad 2 \leqslant i \leqslant I-1, \quad 2 \leqslant j \leqslant J-1,\end{aligned}$$

where, for convenience, we have assumed a rectangular I by J grid; AX and AY are the x- and y-components of $\mathbf{A}$; Δx and Δy are the x- and y-grid intervals (assumed constant); and $\phi_{ij} \equiv \phi(i\,\Delta x, j\,\Delta y)$. Equation (A.2) may be expressed as a recursion relation

$$\begin{aligned}\phi_{i,j+1} = [\Delta y^{-2} + AY_{ij}/(2\Delta y)]^{-1}\{S_{ij} - \phi_{i,j-1}[\Delta y^{-2} - AY_{ij}/(2\Delta y)] \\ - \phi_{ij}[-2(\Delta x^{-2} + \Delta y^{-2}) + B_{ij}] - \phi_{i+1,j}[\Delta x^{-2} + AX_{ij}/(2\Delta x)] \\ - \phi_{i-1,j}[\Delta x^{-2} - AX_{ij}/(2\Delta x)]\}.\end{aligned} \tag{A.3}$$

If we know ϕ-values for the row $j = 2$, we can find all other ϕ-values by recursively imposing Eq. (A.3) and the specified boundary constraints (which determine ϕ-values on row $j = 1$ and columns $i = 1$ and $i = I$). *Thus, the original problem with* $(I - 2) \times (J - 2)$ *unknowns has effectively been reduced to finding the* $I - 2$ *unknowns*[2] $\phi_{i,2}$, $2 \leqslant i \leqslant I - 1$. *The GSM's high computational efficiency is achieved by solving directly for these few unknowns and applying the recursion relation* (A.3) *to determine the remaining unknowns.*

In the GSM procedure, Eq. (A.3) and the boundary constraints are applied recursively (marching "upward" as Eq. (A.3) is applied from row $j = 2$ to row $j = J - 1$), after assuming trial second-row values $\phi^p_{i,2}$, $2 \leqslant i \leqslant I - 1$. The result is a particular solution, ϕ^P, which, due to (usually) erroneous trial values $\phi^p_{i,2}$, will not satisfy the "top" boundary constraints. However, the total solution may

[2] For doubly periodic boundary conditions, there are $2I - 4$ unknowns to be determined.

be expressed as $\phi = \phi^p + \phi^h$. The homogeneous solution ϕ^h satisfies Eq. (A.3) with a null source term ($S_{ij} = O$) and counteracts the error of ϕ^p in satisfying the "top" boundary constraint (e.g., if $\phi_{iJ} = 0$, $2 \leqslant i \leqslant I$, is the top boundary constraint, then $\phi^h_{iJ} = -\phi^p_{iJ}$). while satisfying the homogeneous counterpart of the specified boundary constraints on ϕ everywhere else. Again, the second-row values (of ϕ^h) determine whether the top conditions are satisfied after marching upward. The top conditions depend *linearly* on the second-row values chosen to start the homogeneous recursion, Eq. (A.3) with $S_{ij} = O$. This linear relation may be determined once- and for all if **A**, B are fixed; it is independent of the source term S_{ij}, depending only on the difference operator and class of boundary conditions. It may be determined by performing $I - 2$ sweeps with the homogeneous recursion relation, with the nth sweep being started by $\phi_{2,n+1} = 1$ and $\phi_{2,j} = 0, j \neq n + 1$. The resulting matrix of $I - 2$ top-row vectors is then inverted; once this inverse is determined, sequential problems on an $N \times N$ grid may be solved with $O(N^2)$ operations each; no method can be substantially faster than this.

Unfortunately, the size of problems to which direct GSM application is possible is limited by the available computing precision, and the iterative BIR method described in the text must be used in GSM application to high-resolution problems. (Using double precision IBM arithmetic, the Poisson–Dirichlet problem may be solved directly for N as large as 25, if one performs extra iterations to relax round-off error effects; if previous results are available, these iterations may be reduced by accurate extrapolation of second-row values $\phi^p_{2,j}$.) The reason for this restriction is best illustrated by considering the discrete Poisson equation on a grid with equal spacing in both directions.

$$\nabla^2\phi_{ij} = \phi_{i+1,j} + \phi_{i,j+1} + \phi_{i-1,j} + \phi_{i,j-1} - 4\phi_{ij} = q_{ij}\,. \tag{A.4}$$

As the recursion analogous to Eq. (A.3) is carried forward, round-off error introduces a spurious solution $\tilde{\phi}$, which is the fastest-growing solution to $\nabla^2\tilde{\phi} = 0$ resolved by the discrete grid. The continuous analog is $\tilde{\phi} \propto \sin kx \cdot e^{ky}$. The discrete $\tilde{\phi}$ changes sign from grid point to grid point in the x-direction and increases by a constant factor α from row to row in the y-direction. An equation for α may be derived from Eq. (A.4):

$$\alpha^2 + 6\alpha + 1 = 0.$$

Thus, $\alpha = 3 \pm 8^{1/2}$ with the two roots describing one growing and one decaying solution. The growing solution $\tilde{\phi}$ increases by a factor of 5.83 from row to row, so that an initial round-off error of 10^{-16} grows to 5×10^{-6} after 14 applications of the recursion relation.

Although the GSM generalizes to problems with more than two dimensions, it is most efficient when applied to one- and two-dimensional problems. For an

M-dimensional problem with N grid intervals in each direction, the GSM requires $O(N^{2M-2})$ operations to generate starting values for the homogeneous solution step.

Finally, if $AY = 2/\Delta y$, the recursion relation (A.3) is singular. This type of singularity can be avoided in several ways. One is to increase resolution, so that $|AY| < 2\,\Delta y$ everywhere. Another is to change the direction of GSM sweeps: if AY is sufficiently smooth, reverse the direction of the sweeps; or, sweep in the x-direction if $|AX| < 2/\Delta x$ everywhere. This type of singularity is referred to in Section 5 of the text.

ACKNOWLEDGMENTS

During the first author's recent service at McGill University, this research was supported by the National Research Council of Canada, under grant number 280-98. More recently, it has been supported by the Office of Naval Research. We are also indebted to Drs. Steve Piacsek, Glyn Roberts, Tom Warn, and Andrew Staniforth for valuable comments in the course of this research, and to Ms. Marty Schmidt for typing the manuscript.

REFERENCES

1. D. E. DIETRICH, A Numerical Study of Rotating, Baroclinic Flows in the Rotating Annulus Experiments, Ph.D. Thesis, Florida State University, 1972.
2. D. E. DIETRICH, *J. Meteorol. Soc. Japan* **52** (1974), 115–116.
3. D. R. EDWARDS AND K. F. HANSEN, *Nucl. Sci. Eng.* **25** (1966), 58–65.
4. I. HIROTA, T. TOKIOKA, AND M. NISHIGUCHI, *J. Meteorol. Soc. Japan* **48** (1970), 161–167.
5. M. KWIZAK AND A. ROBERT, *Monthly Weather Rev.* **99** (1971), 32–36.
6. G. J. MCAVANEY AND L. M. LESLIE, *J. Meteorol. Soc. Japan* **50** (1972), 136–137.
7. P. E. MERILEES, *Atmosphere* **11** (1973), 13–20.
8. J. J. O'BRIEN AND H. E. HURLBURT, *J. Phys. Oceanogr.* **2** (1972), 14–26.
9. D. W. PEACEMAN AND H. H. RACHFORD, JR., *J. Solar Indust. Appl. Math.* **3** (1955), 28–41.
10. S. PIACSEK AND G. P. WILLIAMS, *J. Comp. Phys.* **6** (1970), 392–405.
11. P. J. ROACHE, A New Direct Method for the Discretized Poisson Equation, *in* "Proceedings of the Second International Conference on Numerical Methods in Fluid Dynamics, University of California, Berkeley, September 16–19, 1970," pp. 48–53, 1971.
12. G. ROBERTS, personal communication.
13. E. L. WACHSPRESS, "Iterative Solution of Elliptic Systems," Prentice-Hall, Englewood Cliffs, N.J., 1966.
14. T. WARN, personal communication.
15. D. E. DIETRICH, J. Meteorol. Soc. Japan 53 (1975), to appear in June issue.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.