Tellus (1977), *29*, 512–522

On weighted-mean schemes for the finite-difference approximation to the advection-diffusion equation

By MANUEL E. FIADEIRO and GEORGE VERONIS, *Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06520, U.S.A.*

(Manuscript received January 19, 1977)

ABSTRACT

The weighted-mean scheme is a method for constructing finite-difference approximations of second-order partial differential equations of the advection-diffusion type using only the center and adjacent points in each space direction. The scheme tends to a centered-difference formulation for strongly diffusive cases and to an upstream formulation for strongly advective cases. The error of approximation is $O(h^2)$ or better, when h tends to zero, and the scheme assures stability and convergence to all iterative methods no matter how large the grid size. The scheme thus makes it possible to choose the biggest grid size suitable for each specific problem thereby reducing the computing time considerably.

1. Introduction

Numerical solutions to the Navier-Stokes equations are most often obtained by finite-difference methods which involve approximations at several stages. The goal is to derive a solution which is as close as possible to the continuous solution of the system. In principle, it is always possible to achieve this goal with a sufficiently fine grid network, but practical and economic considerations may put the goal out of reach. We shall deal here with the problems generated by the discretization of the space domain and shall propose a method for obtaining accurate results economically. More specifically, we must first ensure that the correct qualitative behavior of the system is achieved and then that the correct quantitative results are optimal for the effort expended.

Our analysis is for the steady advection-diffusion equation of a dynamically passive tracer in an incompressible fluid with a known velocity field. This equation contains many of the essential difficulties of the Navier-Stokes equations but it is simpler to analyse. Since the focus is on the space discretization, the treatment of the steady system is in no sense restrictive; the same spatial problems must be faced for the transient system irrespective of how the time derivatives are treated.

Therefore, consider the equation for the conservation of a substance, C,

$$\nabla \cdot (C\mathbf{v}) = \nabla \cdot ([K]\nabla C) + J$$

or (1)

$$\nabla \cdot (C\mathbf{v} - [K]\nabla C) = J.$$

C is the concentration of the substance, $\mathbf{v} = (u, v, w)$ is the known velocity field, $[K]$ is the diagonal matrix of the known coefficients of eddy diffusion and J (referred to as the consumption term) is the sum of all internal sources or sinks of the substance. J may be a function of C (as for a radioactive tracer) or of position but not of gradients of C. If the substance is being produced (as is oxygen during photosynthesis, for example), J is positive.

The equation for the conservation of fluid mass is given when $C = 1$ and $J = 0$ and reduces to

$$\nabla \cdot \mathbf{v} = 0. \tag{2}$$

2. The matrix for the space domain

Different finite-difference systems for eq. (1) may vary because of the type of grid used (uniform or not) and also because of the methods and the

number of points used to approximate derivatives by differences. In all cases, however, the problem ultimately reduces to a system of equations which can be written symbolically in matrix form as

$$[Q]\mathbf{C} = \mathbf{B}. \tag{3}$$

C is the vector of concentrations, $[Q]$ the matrix of coefficients and the vector, **B**, contains inhomogeneous and boundary terms. Specifically, if we consider a difference scheme which for each point involves only the point and its immediate neighbors, each of the equations will have the form

$$\sum_n Q_n C_n + Q_0 C_0 = B_0 \tag{4}$$

where the summation is taken over the points adjacent to the center point (the latter denoted by subscript 0). The summation will involve two, four, or six points for one-, two-, or three-dimensional problems. Thus, typically, $[Q]$ will be a large-order, sparse, band-diagonal matrix. In these circumstances, iterative methods are usually the best way to obtain the solution of the system of eqs. (3).

The requirements for convergence of the iteration procedure depend upon the method used, but it can be shown that all reasonable iterative methods are stable and convergent if matrix $[Q]$ has the following properties;

(i) No diagonal terms vanish, $Q_0 \neq 0$, for all points.

(ii) The matrix is of positive type, $Q_0/Q_n < 0$, for all neighboring points.

(iii) The matrix is diagonally dominant, $|Q_0| \geqslant \sum_n |Q_n|$, at every point, with strict inequality for at least one point.

(iv) All points are connected, that is, the grid connects all the points in one domain by a sequence of neighbors.

Property (iv) depends on how the grid is laid. If the total domain of the variable is not connected it can be divided into two or more connected domains each with a set of boundary conditions. Properties (i), (ii) and (iv) make the matrix irreducible. Properties (i) and (ii) make it an L-matrix (Young & Gregory, 1973, ch. 16). Therefore, in the procedure proposed below we have satisfied conditions (i) to (iv) even though it is not clear whether they are absolutely necessary.

3. The finite-difference approximations

The derivation of eq. (1) makes use of the divergence theorem applied to a volume the size of which is allowed to shrink to zero. In the finite-difference formulation it pays to backtrack one step and work with surface fluxes across the faces of a finite volume thereby ensuring exact conservation of the substance.

The space domain is divided into a rectangular grid network of points where each point is at the center of a small cell. Without loss of generality we shall assume regular grid intervals of sizes h, l, m in the x, y, z directions respectively. We shall develop a finite-difference formulation of (1) with solutions that agree within $O(h^2,l^2,m^2)$ with the solution of (1) at the grid points for small grid sizes. The matrix of coefficients will obey conditions (i), (ii) and (iii) above for any grid size.

The following notation is used: Values at the generic point (i,j,k) are denoted by the subscript zero, thus, $F(i,j,k)$ is written F_0. Values at other points are identified by the value of the subscript that differs from i, j, or k. Thus, $F(i + 1/2,j,k)$ is written as $F_{i+1/2}$, $F(i,j-1,k)$ as F_{j-1}, etc.

In cartesian coordinates (1) becomes

$$\frac{\partial}{\partial x}\left(uC - K_x\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial y}\left(vC - K_y\frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial z}\left(wC - K_z\frac{\partial C}{\partial z}\right) = J. \tag{5}$$

We can use the divergence theorem to write (5) in term of a surface integral over a small cell of size hlm or, alternatively, we can approximate the derivatives in (5) by finite differences to obtain an expression in terms of fluxes at the interfaces of the cell. The result is:

$$\frac{1}{h}(F_{i+1/2} - F_{i-1/2}) + \frac{1}{l}(F_{j+1/2} - F_{j-1/2}) + \frac{1}{m}(F_{k+1/2} - F_{k-1/2}) = J_0. \tag{6}$$

Here,

$$F_{i\pm 1/2} = \left(uC - K_x\frac{\partial C}{\partial x}\right)_{i\pm 1/2},$$

$$F_{j\pm 1/2} = \left(vC - K_y\frac{\partial C}{\partial y}\right)_{j\pm 1/2},$$

$$F_{k\pm 1/2} = \left(wC - K_z\frac{\partial C}{\partial z}\right)_{k\pm 1/2}.$$

Equation (6) is exact if the fluxes have the mean values over the interfaces and J_0 the mean value of J in the cell. The equation is correct only to $O(h^2,l^2,m^2)$ if F and J are given at mid-point values.

For mass conservation we have $C = 1$ and $J = 0$ so that (6) becomes

$$\frac{1}{h}(u_{i+1/2} - u_{i-1/2}) + \frac{1}{l}(v_{j+1/2} - v_{j-1/2}) + \frac{1}{m}(w_{k+1/2} - w_{k-1/2}) = 0. \tag{7}$$

It is sometimes desirable to satisfy mass conservation with no truncation error. If the velocity is given analytically, eq. (7) can be made exact by averaging the component velocities over the surfaces designated by the subscripts. If the velocity is given as an array of numbers, the array can be adjusted by suitable interpolation to satisfy (7) exactly. If the velocity must be computed as part of the calculation, it is not generally possible to avoid truncation errors in (7).

We now make three statements about the fluxes at half-interval points:

(a) The flux is expressed as a linear combination of the concentrations at the adjacent grid points. Accordingly,

for the concentration:

$$\frac{1}{h}F_{i+1/2} = \beta_{i+1}C_0 - \alpha_{i+1}C_{i+1},$$
$$\frac{1}{h}F_{i-1/2} = \alpha_{i-1}C_{i-1} - \beta_{i-1}C_0,$$
$$\frac{1}{l}F_{j+1/2} = \beta_{j+1}C_0 - \alpha_{j+1}C_{j+1},$$
$$\frac{1}{l}F_{j-1/2} = \alpha_{j-1}C_{j-1} - \beta_{j-1}C_0,$$
$$\frac{1}{m}F_{k+1/2} = \beta_{k+1}C_0 - \alpha_{k+1}C_{k+1},$$
$$\frac{1}{m}F_{k-1/2} = \alpha_{k-1}C_{k-1} - \beta_{k-1}C_0,$$

for fluid continuity

$$\frac{1}{h}u_{i+1/2} = \beta_{i+1} - \alpha_{i+1},$$
$$\frac{1}{h}u_{i-1/2} = \alpha_{i-1} - \beta_{i-1},$$
$$\frac{1}{l}v_{j+1/2} = \beta_{j+1} - \alpha_{j+1},$$
$$\frac{1}{l}v_{j-1/2} = \alpha_{j-1} - \beta_{j-1},$$
$$\frac{1}{m}w_{k+1/2} = \beta_{k+1} - \alpha_{k+1},$$
$$\frac{1}{m}w_{k-1/2} = \alpha_{k-1} - \beta_{k-1}.$$

(8)

Hence, only α or β is independent in each flux equation. Furthermore, from (6), (7) and (8) we have

$$\sum_n \alpha_n = \sum_n \beta_n \tag{9}$$

and

$$\sum_n \alpha_n C_n - \alpha_0 C_0 = -J_0 \tag{10}$$

where the sum is taken over the neighboring points. The value of α_0 is $\sum_n \beta_n$, which is equal to $\sum_n \alpha_n$ to the accuracy of conservation of mass. In practice, it is best to use $\alpha_0 = \sum_n \alpha_n$ because that ensures conservation of C in (10). Thus, even though the individual values of C may still involve an error, the total conservation of C will not.

(b) Since diffusion is a symmetric process, the diffusive flux is evaluated by central differences:

$$K_x \frac{\partial C}{\partial x}\bigg|_{i+1/2} = K_{i+1/2}\frac{C_{i+1} - C_0}{h},$$

$$K_x \frac{\partial C}{\partial x}\bigg|_{i-1/2} = K_{i-1/2}\frac{C_0 - C_{i-1}}{h}, \text{ etc} \tag{11}$$

(c) The concentration in the advective flux term at the half-interval points is taken as a weighted mean of the values of the adjacent points. Thus,

$$C_{i+1/2} = \frac{1-\sigma_{i+1/2}}{2} C_{i+1} + \frac{1+\sigma_{i+1/2}}{2} C_0,$$
$$C_{i-1/2} = \frac{1-\sigma_{i-1/2}}{2} C_0 + \frac{1+\sigma_{i-1/2}}{2} C_{i-1}, \text{ etc.} \tag{12}$$

The values of σ are to be evaluated as functions of h, l, m, and of the values of K and the velocity components at the half-interval points. The sign of σ should take on the sign of the velocity component and in absolute size $|\sigma| \leqslant 1$. We note that $\sigma = +1$ and $\sigma = -1$ correspond to the advection of the upstream value as we would expect when diffusion is negligible. For $\sigma = 0$, $C_{i\pm1/2}$ is the arithmetic mean of $C_{i\pm1}$ and C_0 (the centered mean) as we would expect when diffusion processes dominate.

Now with (11) and (12) in the flux terms and use of (8) at the point $i + 1/2$ we obtain

$$\frac{1}{h}\left\{ u_{i+1/2}\left[\frac{1-\sigma_{i+1/2}}{2} C_{i+1} + \frac{1+\sigma_{i+1/2}}{2} C_0 \right] - K_{i+1/2} \frac{C_{i+1} - C_0}{h} \right\}$$
$$= \left(\alpha_{i+1} + \frac{u_{i+1/2}}{h} \right) C_0 - \alpha_{i+1} C_{i+1} \tag{13}$$

or collecting coefficients of C_{i+1} and C_0,

$$\left[\frac{u_{i+1/2}}{2h}(1-\sigma_{i+1/2}) - \frac{K_{i+1/2}}{h^2} + \alpha_{i+1} \right] \times (C_{i+1} - C_0) = 0. \tag{14}$$

Since C is generally not constant, we have $C_{i+1} \neq C_0$ so that the term in the square brackets must vanish. Therefore,

$$\alpha_{i+1} = \frac{K_{i+1/2}}{h^2} - \frac{u_{i+1/2}}{2h}(1-\sigma_{i+1/2}),$$
$$\beta_{i+1} = \frac{K_{i+1/2}}{h^2} + \frac{u_{i+1/2}}{2h}(1+\sigma_{i+1/2}). \tag{15a}$$

A similar procedure at the other five surfaces leads to

$$\alpha_{i-1} = \frac{K_{i-1/2}}{h^2} + \frac{u_{i-1/2}}{2h}(1+\sigma_{i-1/2}),$$
$$\beta_{i-1} = \frac{K_{i-1/2}}{h^2} - \frac{u_{i-1/2}}{2h}(1-\sigma_{i-1/2}) \tag{15b}$$

$$\alpha_{j+1} = \frac{K_{j+1/2}}{l^2} - \frac{v_{j+1/2}}{2l}(1-\sigma_{j+1/2}),$$
$$\beta_{j+1} = \frac{K_{j+1/2}}{l^2} + \frac{v_{j+1/2}}{2l}(1+\sigma_{j+1/2}) \tag{15c}$$

$$\alpha_{j-1} = \frac{K_{j-1/2}}{l^2} + \frac{v_{j-1/2}}{2l}(1+\sigma_{j-1/2}),$$
$$\beta_{j-1} = \frac{K_{j-1/2}}{l^2} - \frac{v_{j-1/2}}{2l}(1-\sigma_{j-1/2}) \tag{15d}$$

$$\alpha_{k+1} = \frac{K_{k+1/2}}{m^2} - \frac{w_{k+1/2}}{2m}(1-\sigma_{k+1/2}),$$
$$\beta_{k+1} = \frac{K_{k+1/2}}{m^2} + \frac{w_{k+1/2}}{2m}(1+\sigma_{k+1/2}) \tag{15e}$$

$$\alpha_{k-1} = \frac{K_{k-1/2}}{m^2} + \frac{w_{k-1/2}}{2m}(1+\sigma_{k-1/2}),$$
$$\beta_{k-1} = \frac{K_{k-1/2}}{m^2} - \frac{w_{k-1/2}}{2m}(1-\sigma_{k-1/2}) \tag{15f}$$

Substituting (15) in (10) we obtain

$$\left[\frac{K_{i+1/2}}{h^2} - \frac{u_{i+1/2}}{2h}(1-\sigma_{i+1/2}) \right] C_{i+1}$$
$$+ \left[\frac{K_{i-1/2}}{h^2} + \frac{u_{i-1/2}}{2h}(1+\sigma_{i-1/2}) \right] C_{i-1}$$
$$+ \left[\frac{K_{j+1/2}}{l^2} - \frac{v_{i+1/2}}{2l}(1-\sigma_{j+1/2}) \right] C_{j+1}$$
$$+ \left[\frac{K_{j-1/2}}{l^2} + \frac{v_{j-1/2}}{2l}(1+\sigma_{j-1/2}) \right] C_{j-1}$$
$$+ \left[\frac{K_{k+1/2}}{m^2} - \frac{w_{k+1/2}}{2m}(1-\sigma_{k+1/2}) \right] C_{k+1}$$
$$+ \left[\frac{K_{k-1/2}}{m^2} + \frac{w_{k-1/2}}{2m}(1+\sigma_{k-1/2}) \right] C_{k-1}$$
$$- \alpha_0 C_0 = -J_0, \quad \alpha_0 = \sum_n \alpha_n. \tag{16}$$

We must still find the values of σ that lead to the best approximate solutions for the values of C. Consider first the one-dimensional case with $J = 0$ and constant K. Conservation of mass yields u = constant so that $u_{i-1/2} = u_{i+1/2} = u$. Then (16) reduces to

$$\left[\frac{K}{h^2} - (1-\sigma)\frac{u}{2h}\right]C_{i+1} - 2\left(\frac{K}{h^2} + \frac{\sigma u}{2h}\right)C_0 + \left[\frac{K}{h^2} + (1+\sigma)\frac{u}{2h}\right]C_{i-1} = 0. \tag{17}$$

Multiplying by h^2/K leads to

$$[1 - (1-\sigma)\theta]C_{i+1} - 2(1+\sigma\theta)C_0 + [1 + (1+\sigma)\theta]C_{i-1} = 0 \tag{18}$$

where

$$\theta = \frac{uh}{2K}. \tag{19}$$

Equation (18) is the finite-difference approximation to the differential equation

$$K\frac{\partial^2 C}{\partial x^2} - u\frac{\partial C}{\partial x} = 0. \tag{20}$$

Young & Gregory (1973) show that the finite-difference equation (we shall call it the *exact equivalent equation*)

$$e^{-\theta}C_{i+1} - 2\cosh\theta C_0 + e^{\theta}C_{i-1} = 0 \tag{21}$$

has solutions that agree identically at the points x_i with the solutions to the continuous eq. (20). We can reduce (18) to the form (21) with the choice

$$\sigma = \coth\theta - 1/\theta. \tag{22}$$

Thus, (22) eliminates errors of all orders from the finite-difference system.

In Fig. 1 the curve marked 1 shows values of σ vs θ which lead to the exact equivalent equation. σ is an odd function of u and tends to 1 for increasing θ.

For condition (ii) in Section 2 to be obeyed by (18), the coefficient of C_{i+1} (if u is positive) or the coefficient of C_{i-1} (if u is negative) must be greater than zero, that is,

$$1 + \sigma\theta > |\theta|. \tag{23}$$

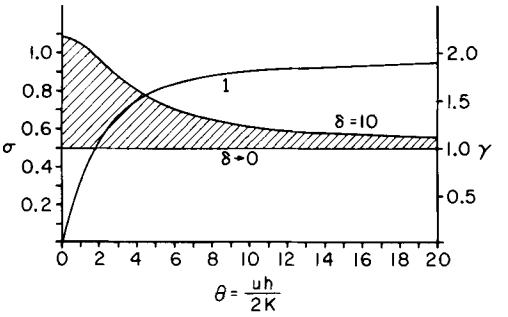

Fig. 1. Values of σ vs θ for the exact-equivalent equation are shown by the curve marked 1. Values of γ vs σ and δ are shown in the shaded portion of the region. The upper bounding curve corresponds to $\delta \gg 1$ and the lower to $\delta \ll 1$. Intermediate values of δ lead to points in the shaded region for each value of θ.

For $|\theta| \leqslant 1$ or $h \leqslant 2K/u$, the condition is satisfied with centered-differences ($\sigma = 0$). This is, in fact, the limit for the grid size that can be used to solve the finite-difference equations with the centered-difference approximation. For $|\theta| > 1$, $|\sigma|$ should be greater than $1 - 1/|\theta|$. Equation (22) is thus consistent with the requirement for a positive matrix. In the limit of large $|\theta|$, (22) reduces essentially to $1 - 1/|\theta|$. The system then becomes purely advective. The best way to solve the system in this case is along characteristics.

When the centered-difference scheme is used ($\sigma = 0$) the coefficient of C_0 in (17) is $-2K/h^2$ where K is the diffusion coefficient. In the more general formula (17) the coefficient of C_0 is $2K(1 + \sigma\theta)/h^2$. Hence, for non-vanishing σ, the weighted-mean scheme can be viewed as a centered-difference scheme with a virtual diffusion coefficient given by $K(1 + \sigma\theta) = K(\theta/\tanh\theta)$. In the limit of small h (or θ), σ tends to $\theta/3$ and the virtual diffusion coefficient becomes $K[1 + (u^2h^2/12K^2)]$, where the term $u^2h^2/12K^2$ ($= \theta^2/3$) is necessary to correct errors of $O(h^2)$ introduced by the centered-difference scheme. More generally, the factor $\theta/\tanh\theta$ corrects errors of all orders.

In more complex situations, ones with variable coefficients and/or more than one dimension, there is no general technique for obtaining an exact equivalent equation. Indeed, it is not generally possible even to eliminate errors of $O(h^2,l^2,m^2)$, although higher-order approximations to the derivatives can be used to reduce second-order errors. One can apply the results of the present section to those problems, treating the system

locally as if the one-dimensional constant-coefficient analysis were applicable in each direction. The value of σ will change from point to point since it is dependent on the respective values of K, u, v, w, h, l, m. Accordingly, the coefficients of (16) will have the following values:

$$\begin{aligned}
\alpha_{i+1} &= \frac{u_{i+1/2}}{2h}\left[\coth\left(\frac{hu_{i+1/2}}{2K_{i+1/2}}\right) - 1\right] \\
\alpha_{i-1} &= \frac{u_{i-1/2}}{2h}\left[\coth\left(\frac{hu_{i-1/2}}{2K_{i-1/2}}\right) + 1\right] \\
\alpha_{j+1} &= \frac{v_{j+1/2}}{2l}\left[\coth\left(\frac{lv_{j+1/2}}{2K_{j+1/2}}\right) - 1\right] \\
\alpha_{j-1} &= \frac{v_{j-1/2}}{2l}\left[\coth\left(\frac{lv_{j-1/2}}{2K_{j-1/2}}\right) + 1\right] \\
\alpha_{k+1} &= \frac{w_{k+1/2}}{2m}\left[\coth\left(\frac{mw_{k+1/2}}{2K_{k+1/2}}\right) - 1\right] \\
\alpha_{k-1} &= \frac{w_{k-1/2}}{2m}\left[\coth\left(\frac{mw_{k-1/2}}{2K_{k-1/2}}\right) + 1\right] \\
\alpha_0 &= \sum_n \alpha_n.
\end{aligned} \tag{24}$$

The formulation uses only a five- and seven-point operator for two and three dimensions respectively and is antisymmetric in relation to the velocity field. When the components change sign, the coefficients upstream and downstream of the point are automatically reversed, a feature particularly useful in computer programming because the sign of the velocity components need not be known *a priori*.

4. The consumption term J

As long as the consumption term does not involve gradients of C or a positive first-order rate it does not affect the diagonal dominance of the coefficient matrix or the numerical stability of the procedure. However, the adopted form affects the accuracy of the solution. Suppose, for example, that J is due to radioactive decay so that

$$J = -\lambda C. \tag{25}$$

The most direct way to treat this is to write it as

$$\lambda C = \lambda C_0 \tag{26}$$

in which case eq. (18) becomes

$$\begin{aligned}
&[1 - (1 - \sigma)\theta]C_{i+1} - 2(1 + \sigma\theta + \delta/2)C_0 \\
&\quad + [1 + (1 + \sigma)\theta]C_{i-1} = 0
\end{aligned} \tag{27}$$

where

$$\delta = \lambda h^2/K. \tag{28}$$

The exact-equivalent equation in this case is

$$e^{-\theta}C_{i+1} - 2\cosh\sqrt{\theta^2 + \delta}\,C_0 + e^{\theta}C_{i-1} = 0. \tag{29}$$

and it is easily seen that no choice of σ can reduce (27) to (29). Now σ was introduced to give a weight to the advection term *vis a vis* the diffusion term and we saw that it introduced a virtual diffusion into the finite-difference system. Since it does not provide sufficient flexibility to yield the exact-equivalent equation, we introduce another parameter, one associated with radioactive decay, and write

$$\lambda C = \gamma\lambda C_0. \tag{30}$$

Then (27) becomes

$$\begin{aligned}
&[1 - (1 - \sigma)\theta]C_{i+1} \\
&\quad - 2\left(1 + \sigma\theta + \frac{\gamma\delta}{2}\right)C_0 \\
&\quad + [1 + (1 + \sigma)\theta]C_{i-1} = 0.
\end{aligned} \tag{31}$$

Equation (31) can now be made identical to (29) if σ and γ are chosen as

$$\begin{aligned}
\sigma &= \coth\theta - 1/\theta, \\
\gamma &= 2\theta(\cosh\sqrt{\theta^2 + \delta} - \cosh\theta)/\delta\sinh\theta.
\end{aligned} \tag{32}$$

Accordingly, σ has the same value as for the case without decay. We note that in addition to a virtual diffusion the present system involves a virtual radioactive decay coefficient of magnitude $\gamma\lambda$ instead of λ. For small increments, i.e. h (hence θ) $\to 0$, we obtain $\gamma \to 1$, and the usual form for the radioactive decay is recovered.

Values of γ vs σ and δ are shown in the shaded portion of Fig. 1. For small λ (hence δ) γ is essentially 1 but for large λ, γ is larger. In

particular, for weak advection ($\theta \ll 1$) and strong decay ($\delta \gg 1$) γ has a value as high as 2.16. The reason for this is that in the absence of advection radioactive decay dominates the local behavior and can lead to relatively sharp gradients. In this case the diffusion process is not well represented by a centered difference and the proper correction is introduced by the enhanced decay. As θ increases, advection has a larger effect so that a smaller correction for radioactive decay is required.

In the general case with variable coefficients or multiple dimensions we can follow the same procedure as outlined in Section 3. The term α_0 now becomes

$$\alpha_0 = \sum \alpha_n + \gamma_0 \lambda.$$

The values of α_n are still evaluated by (24) since they are unaffected by the presence of λ. For most practical cases radioactive decay does not dominate the local behavior and γ_0 would not be much different from unity. In the most extreme cases where γ_0 could exceed the value 2 in one of the three directions an approximate treatment would be to choose

$$\gamma_0 = \tfrac{1}{3}(\gamma_1 + \gamma_2 + \gamma_3)$$

where the γ_i are the values obtained from (32) for each of the directions treated separately.

5. A test calculation

The procedure proposed at the end of Section 3 was tested in a simple, two-dimensional case for the square region shown in Fig. 2. The velocity field is described by $u = U(1 - x/L)$, $v = Uy/L$. The stream function, ψ, defined by $u = -\partial\psi/\partial y$, $v = \partial\psi/\partial x$, is given by $\psi = Uy(x/L - 1)$. Hence, the streamlines are the hyperbolas shown in Fig. 2 with flow entering along the left boundary and leaving through the upper boundary. Boundary conditions are $C = 0$ at left and bottom and $C = 1$ at right and top. In the purely advective case the concentration vanishes throughout the region. In the absence of advection ($U = 0$) C varies smoothly from the value $C = 0$ at the lower left corner to $C = 1$ at the upper right corner.

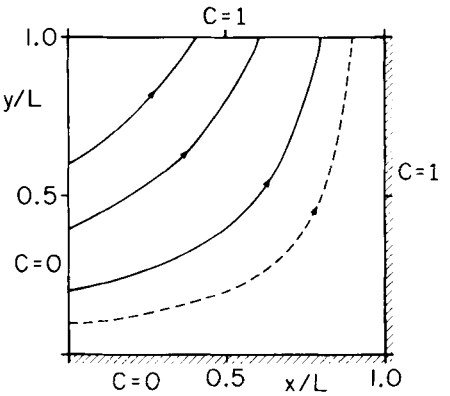

Fig. 2. The flow pattern for the first test calculation is shown as a set of hyperbolae. The fixed concentrations at the boundaries are $C = 0$ along left and bottom and $C = 1$ along top and right, as shown.

With the x and y coordinates scaled by L and u and v by U the advection-diffusion equation reduces to

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) C - P\left[\frac{\partial}{\partial x}(uC) + \frac{\partial}{\partial y}(vC)\right] = 0$$

where x, y, u and v are non-dimensional. $P(= UL/K)$ is the Peclét number and serves as a measure of the intensity of advection relative to diffusion. Although P is the only physical parameter in the continuous system, the significant parameter for the finite-difference system is $\Theta = Uh/2K = P/2N$ where $N = L/h$ is the number of grid intervals of size h in each direction of the square region. (Note that Θ is defined in terms of overall magnitudes in contrast to θ which was defined locally. For the one-dimensional problem discussed in Section 3, θ is constant and equal to Θ.)

The parameter, Θ, serves a dual purpose. It is a measure of the relative intensity of advection via the ratio, U/K, and it also reflects the density of the grid network via the increment, $h = L/N$. Thus, a small value of Θ may correspond either to a system with weak advection or to one with a large number of gridpoints. It is the latter that ensures high accuracy if large gradients are present.

In Table 1 we list values of P and N and the corresponding values of Θ for which calculations were made. Thus, we explored a parametric range over which the effect of advection ranges from dominant (large P) to comparable ($P \simeq 1$) to that of diffusion, and over which the grid network ranges from relatively coarse ($N = 8$) to fine ($N = 64$).

For each combination of P and N calculations were made with the centered-difference scheme ($\sigma = 0$) and with the weighted-mean scheme in which we used eqs. (24) for the value of α. The

Table 1. *Values of* Θ *for the two-dimensional calculation*

N	P				
	128	64	32	16	1.6
8	8	4	2	1	0.1
16	4	2	1	0.5	0.05
32	2	1	0.5	0.25	0.025
64	1	0.5	0.25	0.125	0.0125

solutions were obtained by successive over-relaxation by lines, alternating the x and y directions.

The iterative procedure diverged for the centered-difference equations with $\Theta > 2$. For $\Theta = 2$ convergent solutions were obtained but some local values differed from the correct values by an order of magnitude. With $\Theta = 1$, the maximum error was 15%. Smaller values of Θ led to correspondingly smaller errors in the interior of the system. In general, decreasing h by a factor of two decreased the errors by a factor of four as one would expect for the centered-difference scheme where the errors are $O(h^2)$.

The weighted-mean scheme led to convergent iterations for all values of Θ. For the most strongly advective case ($P = 128$, $N = 8$ and $\Theta = 8$) maximum errors of 25% occurred near the upper left and lower right corners. The reason for these errors is that $N = 8$ provides poor resolution of the large gradients in the relatively thin boundary layers that exist near $x = L$ and $y = L$ for a strongly advective flow. The errors elsewhere in the region did not exceed 6%. Halving the grid increment led to a fourfold decrease of the errors indicating that the errors are of $O(h^2)$. Also, the errors did not exceed those of the centered-difference scheme for any of the cases. Therefore, we conclude that the method is correct to $O(h^2)$, as is the centered-difference procedure, but errors of $O(h^2)$ are quantitatively smaller for the weighted-mean scheme.

The most significant result is that convergent solutions can be obtained for essentially all grid intervals and all velocity amplitudes. With the centered-difference scheme, larger velocities can be treated only if the grid interval is made correspondingly smaller. This restriction often puts the calculation out of reach economically.

Of course, even though one can obtain a numerical solution, the latter may not be accurate, and one is still faced with the problem of interpreting the results. If the grid is too coarse to resolve large gradients in small regions, the numerical values may overshoot or undershoot in those regions. Other inaccuracies may also be present. One can always get around these by decreasing the grid interval sufficiently but normally it is necessary to compromise between really satisfactory accuracy and economic feasibility.

We also tested two other weighted-mean schemes that have been used earlier. Kuo & Veronis (1973) ran a set of numerical experiments on (27) to arrive at an empirical set of values of σ vs θ which can be approximated very closely by $\sigma = \tanh \theta/3$. For small θ this relation reduces to $\sigma = \theta/3$ in agreement with ours. For larger θ their values for σ exceed the ones obtained here by as much as 15%. Calculations with $\sigma = \tanh \theta/3$ for $\Theta = 8$, yield errors at least twice as large as those with the scheme proposed in this paper. As Θ is decreased, the results of the two methods merge.

The form used by Fiadeiro (1975) is $\sigma = (-1 + \sqrt{1 + 4\theta^2})/2\theta$, which tends to θ for small θ and is also larger than our σ. Calculations for the two-dimensional problem using this form led to errors larger than those of either of the other two methods. However, even this form yielded convergent solutions for all values of Θ.

Another procedure which has often been used in advection-diffusion problems is to evaluate the derivatives in the advection term by upstream differences. We did not test this scheme specifically because it has errors of $O(h)$.

6. A second test calculation

Our second test calculation involves determination of the temperature field in a two-dimensional, square, Bénard convection cell for a liquid layer heated uniformly from below and cooled from above. The system is cyclic in the horizontal (x) direction and the boundaries at the bottom and top of the layer ($z = 0, \pi L$) are stress-free, perfect conductors. The cell has width and height πL. The space dimensions are made non-dimensional by use of the length scale, L. The velocity has maximum amplitude, U, which is used in the non-dimensionalization.

The non-dimensional temperature equation has the form

$$P\left[\frac{\partial}{\partial x}(uT) + \frac{\partial}{\partial z}(wT)\right] = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}\right)T$$

with the Peclét number defined as $P = UL/K$ where K is thermal conductivity. The dimensionless velocity field is given by

$$u = -\sin x \cos z, \quad w = \cos x \sin z$$

and satisfies the boundary conditions

$$u = 0 \text{ at } x = 0, \pi; \quad w = 0 \text{ at } z = 0, \pi.$$

Boundary conditions on T are

$$T = 1 \text{ at } z = 0; \quad T = 0 \text{ at } z = \pi;$$
$$\partial T/\partial x = 0 \text{ at } x = 0, \pi.$$

This problem is only part of the free convection problem since the velocity field is given. The full problem has been discussed extensively in the literature (Moore & Weiss, 1973, present an excellent, comprehensive summary). We cannot compare either our calculations or our results directly with published data because we have chosen the velocity field and are calculating only the temperature distribution. Normally, both fields are calculated. Although our chosen velocity field is close to the calculated ones, it is not exactly the same and will lead to somewhat different results for T. The purpose here is to show the advantages of our proposed technique for the numerical solution.

The centered-difference scheme that is normally used to solve this problem has two disadvantages when P is large. First, N, the number of grid intervals, must be large ($> \pi P/8$), if a solution is to be obtained. Therefore, it is not possible to obtain a rough estimate with a crude grid. Second, when P is large and N is sufficiently large to yield a convergent solution, the calculation becomes time-consuming and can be prohibitively expensive. Typically, the larger the value of P, the larger the computational effort.

We used the weighted-mean scheme to obtain the temperature distribution for a range of values of P by relaxation by lines. Initially the temperature was set to zero everywhere except along the $z = 0$ boundary where $T = 1$. Values of Θ ($= Uh/2K$) are shown in Table 2 for $P = 1, 3, 10, 50, 100, 400$ and $N = 16, 32, 64, 128$, when N is the number of grid increments in *each* direction.

Table 2. *Values of Θ for different values of N and P for the Bénard convection calculation*

P	N			
	16	32	64	128
1	.098	.049	.025	.012
3	.29	.145	.073	.037
10	.98	.49	.245	.123
50	4.9	2.45	1.23	.61
100	9.8	4.9	2.45	1.23
400	39.2	19.6	9.82	4.91

Table 3. *Values of Nu for selected values of N and P*

P	N			
	16	32	64	128
1	1.117	—	—	—
3	1.738	—	—	—
10	3.380	3.406	—	—
50	7.122	7.424	7.535	—
100	9.727	10.26	10.56	—
400	—	19.195	20.33	20.99

The first, and perhaps most informative, piece of information about the weighted-mean scheme is that the larger the value of P, the faster the convergence. For example, with $N = 16$ and for $P = 1, 3, 10, 50$ the number of iterations required for convergence was 66, 52, 32 and 15 respectively. Since the scheme reduces to the centered-difference approximation when the velocity is very small ($\Theta \ll 1$), faster convergence for larger P means that the slowest calculations take the same time as the fastest calculations by the centered-difference method. Thus, the present scheme provides access to calculations that cannot be obtained with reasonable effort by the centered-difference scheme.

The Nusselt number, Nu, is the total (conductive plus convective) vertical heat flux averaged across the cell divided by the heat flux that would occur by diffusion alone. In the steady problem Nu must be the same at each level. It was calculated here at all levels as a check on the convergence of the procedure and the difference of maximum to minimum values was always less than 0.01% of the mean.

Values of Nu vs P are shown in Table 3 for different-sized grids. As we stated earlier, we

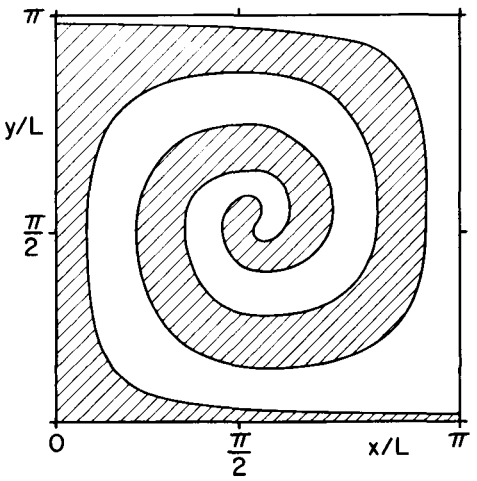

Fig. 3. The spiraling curve is the isotherm $T = 0.5$ for the Bénard convection cell with $P = 400$. The shaded region corresponds to $T > 0.5$ and the clear region to $T < 0.5$. The intertwining tongues indicate the strong effect of convection. (The vertical coordinate is z/L rather than y/L.)

cannot compare our results directly with published ones because the velocity field is taken as given here. However, the value of Nu for $P = 400$ is comparable to that corresponding to the most convective case calculated by Moore & Weiss (1973). With the 128 × 128 grid the present calculation required 44 iterations and five minutes on an IBM 370/158. Moore and Weiss, using a time-stepping, centered-difference method to calculate both the temperature and the velocity fields, required 5 hours on an IBM 360/44 with a grid that was somewhat finer than the one that we have used. (Essentially, their grid interval had to be sufficiently small to satisfy the criterion $\theta_{max} \leq 4$ for convergence.)

One of the advantages of the weighted-mean scheme is that it is possible to obtain results with a coarser grid. With $N = 64$ the heat flux is within 3% of that obtained with $N = 128$. Only 20 iterations and 30 seconds of computing time were required for the calculation with $N = 64$. Thus, one can do the cruder calculation with the coarse grid to bring the system to the neighborhood of the solution and then to interpolate the field onto a finer grid for a more accurate calculation. This procedure is not possible with the centered-difference approximation since the cruder grid does not lead to a convergent system.

A final point that we would like to bring out about the strongly convective cases ($P = 400$) is illustrated in Fig. 3 which shows the contours of the mid-temperature ($T = 0.5$) of the fluid. Strong convection causes a warm ($T > 0.5$, shown shaded) blob of fluid to well up along the left of the cell and then to intertwine with the symmetrically shaped cold ($T < 0.5$) blob of fluid sinking at the right. The bulk of the fluid near the center of the cell is nearly isothermal with $T \simeq 0.5$ but the small deviations from $T = 0.5$ provide a striking demonstration of the effects of strong convection.

7. Summary and conclusions

The development in this paper started from the advection-diffusion equation with the advection terms written in divergent form in order to ensure that conservation integrals are automatically satisfied when the continuous system is approximated by finite differences. Even with this form, however, the centered-difference approximation leads to a system of equations which can be solved by iterative methods only if the parameter uh/k is of $O(1)$. When advection dominates, this restriction on h may be economically impractical, especially for multi-dimensional systems.

Therefore, we have proposed a weighted-mean scheme, in which the advective derivatives at half-intervals are evaluated with a stronger weight on the upstream value. For one-dimensional systems with constant coefficients the weight can be chosen so that the resulting finite-difference equation (which we call the exact equivalent equation) yields solutions at the gridpoints that agree exactly with the values obtained from the continuous system. The finite-difference equation with the weighted-mean scheme contains a term that can be interpreted as a virtual diffusion coefficient that differs from the physical diffusion coefficient by an amount dependent on the velocity and on the size of the grid interval chosen. *It is important to observe that the virtual diffusion coefficient counteracts the distortion introduced by the finite-difference approximation. In this sense it serves to preserve the physical characteristics of the continuous system.*

The weighted-mean scheme has the practical advantage of yielding solutions for relatively large values of the grid size and is, therefore, economical with computer time. Use of it is independent of the sign of the velocity since the weight is an odd

function of the velocity. The method can be used in time-dependent as well as in steady problems.

In multi-dimensional problems, an exact-equivalent equation does not exist in general. We have proposed a procedure in which the weight for the one-dimensional, constant coefficient case is applied locally in each direction. Errors are unavoidable with this procedure but a test calculation indicates that the errors are second order and that they are quantitatively smaller than those obtained with centered differences when the latter can be used. The principal advantage of the proposed technique is that iterative methods are stable and convergent even when very large grid increments are used. However, very large grid increments will introduce distortions and the physical interpretation of the results may not be straightforward. Some experimentation with grid sizes will be necessary in different problems in order that a proper balance between economy and accuracy be obtained.

In a second test calculation we obtained the temperature field in a Bénard convection cell. Our goal here was to test the relative usefulness and accuracy of the method as a function of the intensity of the velocity field and the size of the grid interval. We found that the weighted-mean scheme converges faster for more convective flows. This result is in contrast to that obtained by centered differences where convergence is more elusive as convection increases. Hence, the weighted-mean scheme makes previously inaccessible calculations relatively easy. It also enables one to obtain a crude approximation to a system by using a relatively coarse grid. One can use this crude solution to construct a good, initial-guess solution for a finer grid.

In our treatment here we have focused on the practical and economical features of this scheme. Although we have made use of rigorous analysis to justify the procedure for one-dimensional, constant-coefficient cases, we have not explored the mathematical aspects of more complicated situations. Our hope is that numerical analysts will provide a firmer basis for the use of this method or one that is comparably efficient with computer time. In the meantime, we can certainly recommend the method as one with distinct advantages over the centered-difference scheme that is currently in use.

8. Acknowledgements

This work was supported by NSF Grants DES 73-00424 A01 and OCE76—22141.

REFERENCES

Fiadeiro, M. E. 1975. *Numerical modeling of tracer distributions in the deep Pacific Ocean.* Ph.D. Thesis, UCSD, La Jolla, 226 pp.

Kuo, H.-H. & Veronis, G. 1973. The use of oxygen as a test for an abyssal circulation model. *Deep Sea Res. 20*, 871–888.

Moore, D. R. & Weiss, N. O. 1973. Two-dimensional Reyleigh-Bénard convection. *J. Fluid Mech. 58*, 289–312.

Young, D. & Gregory, R. T. 1973. *A survey of numerical mathematics*, Vol. II. Reading, Mass.: Addison-Wesley Publishing Co., 1099 pp.

О СРЕДНЕВЗВЕШЕННОЙ СХЕМЕ ДЛЯ КОНЕЧНОРАЗНОСТНОЙ АППРОКСИМАЦИИ УРАВНЕНИЯ АДВЕКЦИИ-ДИФФУЗИИ

Средневзвешенная схема является методом построения конечноразностных аппроксимаций для дифференциального уравнения в частных производных второго порядка типа уравнения адвекции-диффузии, использующим только центральную и соседние точки в каждом направлении по пространству. Построенная схема приближается к схеме центральных разностей для случаев строгой диффузии и к схеме односторонних разностей, направленных против потока, для случаев строгой адвекции. Ошибка аппроксимации составляет $O(h^2)$ или лучше, когда h стремится к нулю. Схема обеспечивает устойчивость и сходимость для всех итерационных методов независимо от шага сетки. Таким образом, схема позволяет выбрать наибольший шаг сетки, допустимый для каждой конкретной задачи, тем самым значительно уменьшая время вычислений.