Advective transport via error minimization: enforcing constraints of non-negativity and conservation

W. C. Thacker

National Oceanic and Atmospheric Administration, Atlantic Oceanographic and Meteorological Laboratory, 4301 Rickenbacker Causeway, Miami, Fl 33149 USA

D. Eppel and J. Häuser

Institut für Physik, GKSS Forschungszentrum Geesthacht GmbH, 2054 Geesthacht, West Germany

(Received March 1986; revised June 1986)

An operations research method is presented for deriving a conservative, non-negative computational scheme for advective transport. Finite elements in space and time are used to approximate the solution, and the integral of the square of the residual is minimized over the entire spatial domain and over a single temporal element. Negative values are excluded by inequality constraints and conservation is enforced by Lagrange multipliers. The method is then generalized to show how negative values arising in conventional finite-difference methods can be eliminated.

Keywords: operations research, non-negativity, conservation, finite elements, error-minimization

Advective transport, the process by which anything is carried along by a moving fluid, has two important properties: first, the total amount of whatever is transported remains the same; and second, regardless of how they are redistributed by the flow, positive quantities, such as concentrations of chemical pollutants, must remain positive. It would be highly desirable if numerical simulations of advective transport would also have these properties, but often this is not the case.

Truncation errors necessarily cause the simulations to differ from the exact solutions. For example, the so-called upwind scheme[1] is both conservative and positive, but it is also strongly diffusive; unless very fine grids are used, computed distributions spread out until they hardly resemble the exact solutions. On the other hand, the Leith–Lax–Wendroff scheme[2–4] is only weakly diffusive. It is characterized by the appearance of unwanted ripples, due to numerical dispersion, which plague many computational schemes. An averaging process can be used to smooth out the ripples, but this amounts to increasing the level of diffusion. The problem of ripples versus diffusion has been a subject of much study; for recent examples, see references 5 and 6.

Another form of truncation error, which is found in time-implicit schemes, is nonlocality. Because of the algebraic coupling of the solution at two adjacent time levels, what happens at a single point affects the solution far away. So long as resolution is adequate, the magnitude of remote effects is small. For the purposes of this paper, nonlocality is considered to be acceptable, and attention will be focused on the question of nonphysical negative values.

The negative values are simply a special case of the more general problem of ripples caused by dispersive

0307-904X/86/06438-07/$03.00

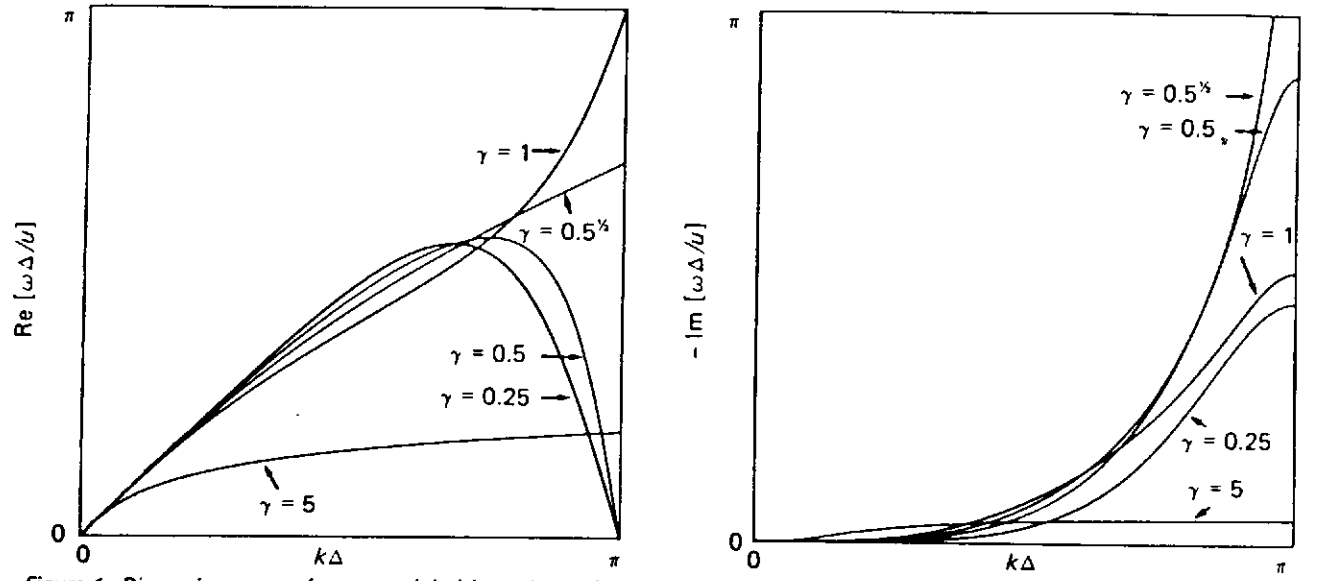

Figure 1 Dispersion curves for error-minimizing scheme for various Courant numbers. Departure of real part of frequency from linearity indicates that poorly resolved waves disperse; presence of negative imaginary part of frequency indicates that poorly resolved waves are damped. This scheme is stable for all Courant numbers

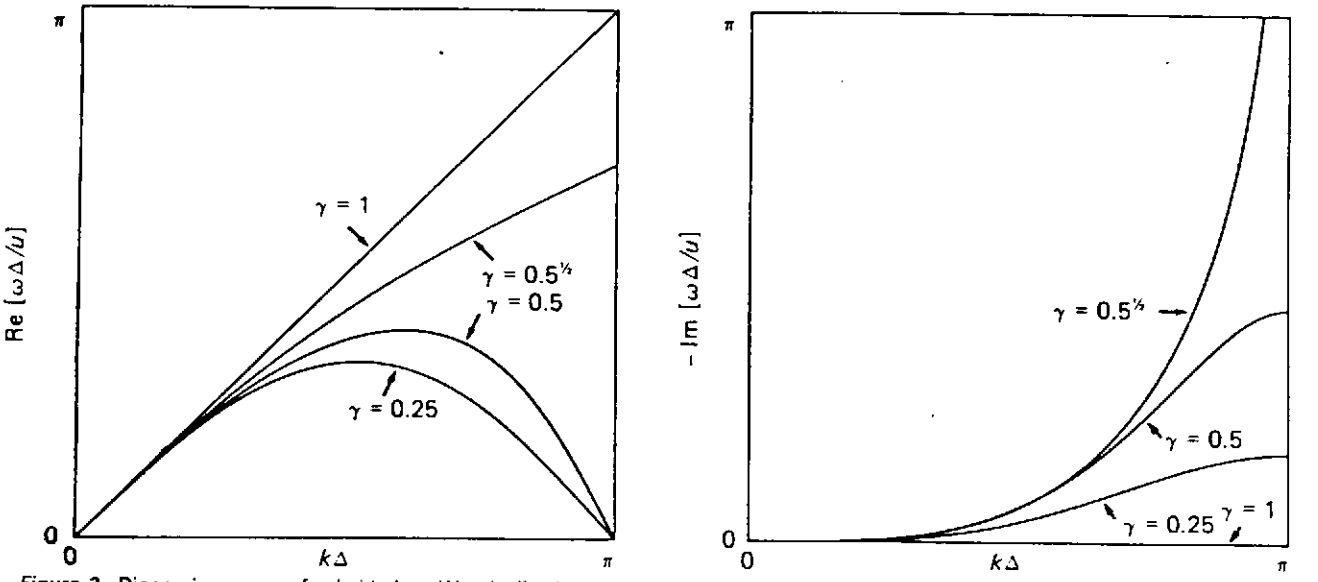

Figure 2 Dispersion curves for Leith–Lax–Wendroff scheme for various Courant numbers. Curves for $\gamma = (0.5)^{\frac{1}{2}}$ are identical with those for error-minimizing scheme; those for $\gamma = 1$ are identical with those for partial-differential equation. Scheme is unstable for $\gamma^2 > 1$

errors. It is possible to address the more general problem by exploiting the monotonicity property of advective transport, i.e. that no new maxima or minima should appear as the distribution evolves. This has been done by a method called flux correction,[7,8] which involves the combination of a higher order scheme leading to oscillations and to a lower order scheme that is strongly diffusive. Flux correction is similar to local smoothing in that local minima are filled-in from adjacent maxima. Preliminary results indicate that the method presented here, within the restricted context of negativity, can be generalized to address the more general question of monotonicity. Such a generalization, however, is beyond the scope of this paper.

The intent of this paper is to follow the consequences of three simple assumptions to obtain a scheme for computing advective transport. The first assumption is that the scheme should minimize some measure of computational error. The other two are that the scheme be non-negative and conservative. Because of the reasonable nature of these assumptions, the outcome should be interesting and instructive.

Derivation of computational scheme

In order to focus on the properties of positivity and conservation, it is sufficient to restrict attention to the simple case of transport by one-dimensional, steady, uniform flow. If the concentration of advected material is represented by $C(x, t)$, where x and t are spatial and temporal coordinates, respectively, and if u is the constant advecting velocity, then the evolution of C is governed by:

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} = 0 \qquad (1)$$

At the inflow boundary, the value of C should be specified, whilst no boundary condition is required at the

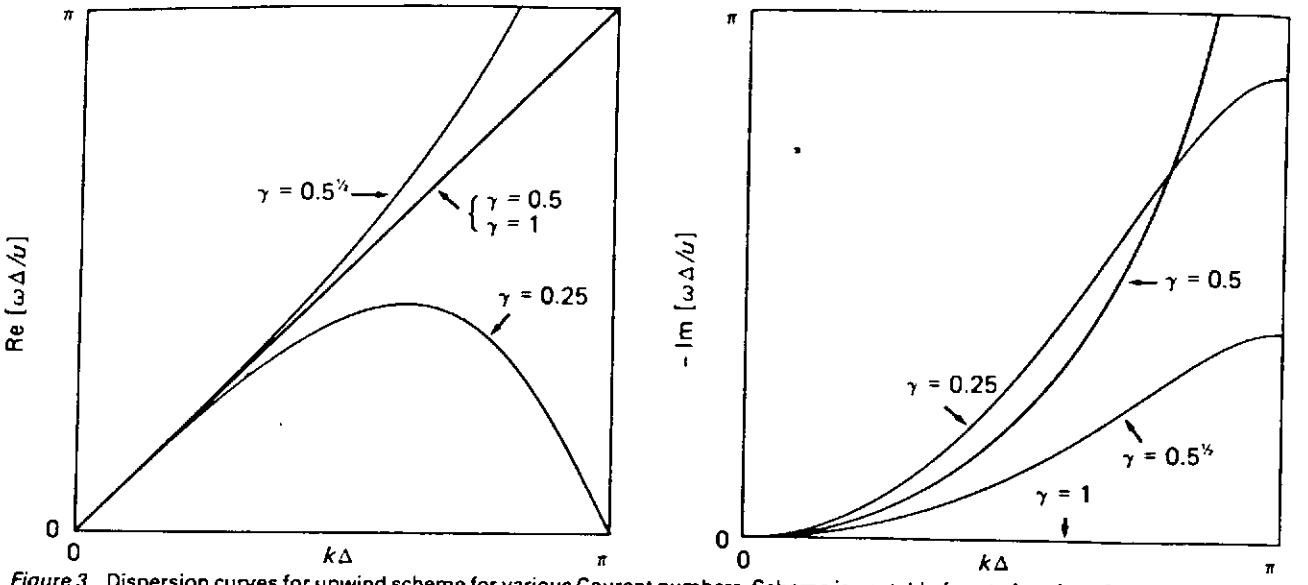

Figure 3 Dispersion curves for upwind scheme for various Courant numbers. Scheme is unstable for $\gamma > 1$ and $\gamma < 0$

outflow boundary. For simplicity, periodic boundary conditions are assumed. When an initial distribution is given, the solution of equation (1) at later times is simply a uniform translation of the initial distribution. (In two or three dimensions, constant flow generalizes to incompressible flow and the distribution remains constant relative to the moving fluid.)

If the solutions to equation (1) are approximated by a function $\tilde{c}(x, t)$ that is piecewise linear in both space and time, then the residual:

$$R(x,t) \equiv \left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x}\right)\tilde{c}(x,t) \tag{2}$$

represents the truncation error. An error density can be defined as any positive function of the residual that vanishes whenever the residual vanishes. Here, it is taken to be the square of the residual, because this leads to the least complicated computational scheme. An error functional is defined as an integral of the error density over the entire spatial domain and over a single step forward in time. The 'best' piecewise bilinear approximation to equation (1) then corresponds to the minimum of this error functional.

The choice of a piecewise bilinear approximation is not essential to this approach. For example, sinusoids might be used to characterize both spatial and temporal variations. What is essential is that the residual resulting from the approximation be used to construct an error functional. Although this approach of minimizing an error functional is quite general, it has not been widely used for time-dependent problems. Zienkiewicz[9] briefly mentions this possibility.

Within the context of this paper, the advantage of this approach is that it allows for a systematic and mathematically well-defined treatment of the positivity constraint. Because the solution is determined by minimizing an error functional, the search for the minimum can be restricted to non-negative functions. A second advantage is that conservation can be guaranteed through the use of a Lagrange multiplier.

If the constraints of positivity and conservation are disregarded, then the minimum of the error functional is determined by the condition that its derivatives vanish. The error functional is an integral over a single temporal element, corresponding to a single step forward in time. Since the solution is known at the beginning of the time-step, only the values at the end of the time-step should be allowed to vary. Once the minimum is found, the process can be repeated to advance step-by-step in time.

If c_j^{n+1} represents the value of the piecewise bilinear approximating function at $x = j\Delta$ and $t = (n+1)\tau$, then the vanishing of the derivative of the error functional with respect to c_j^{n+1} yields the finite difference equation:

$$\frac{1}{\tau}\left[\frac{1}{6}(c_{j+1}^{n+1} + 4c_j^{n+1} + c_{j-1}^{n+1}) - \frac{1}{6}(c_{j+1}^{n} + 4c_j^{n} + c_{j-1}^{n})\right] + \frac{u}{2\Delta}(c_{j+1}^{n} - c_{j-1}^{n})$$
$$= \frac{u^2\tau}{2}\left[\frac{2}{3\Delta^2}(c_{j+1}^{n+1} - 2c_j^{n+1} + c_{j-1}^{n+1}) + \frac{1}{3\Delta^2}(c_{j+1}^{n} - 2c_j^{n} + c_{j-1}^{n})\right] \tag{3}$$

This can readily be seen to be a time-implicit generalization of the Leith–Lax–Wendroff scheme:

$$\frac{1}{\tau}(c_j^{n+1} - c_j^{n}) + \frac{u}{2\Delta}(c_{j+1}^{n} - c_{j-1}^{n}) = \frac{u^2\tau}{2}\left[\frac{1}{\Delta^2}(c_{j+1}^{n} - 2c_j^{n} + c_{j-1}^{n})\right] \tag{4}$$

Both contain discrete analogs of the partial derivatives in equation (1) and 'artificial diffusion' terms that vanish in the limit of infinite resolution. (Although these terms seem to model a diffusive process, both schemes are actually only weakly diffusive.) The differences between equations (3) and (4) can all be attributed to the spatial and temporal averages that are characteristic of finite element methods. For the special case of $u\tau/\Delta = \pm(0.5)^{\frac{1}{2}}$, the two schemes are, in fact, identical.

Both of these schemes are conservative, subject to ripples, and, thus, to nonphysical negative values.

Since the upwind scheme:

$$\frac{1}{\tau}(c_j^{n+1} - c_j^n) + \frac{u}{\Delta}(c_j^n - c_{j-1}^n) = 0 \tag{5}$$

is non-negative, it will be used for comparison. It is conservative but highly diffusive.

The nature of the dispersive and diffusive errors of the three schemes is revealed by a standard Fourier analysis. As these are linear equations with constant coefficients, the solutions can be written as superpositions of complex exponentials, $\exp[i(kx - \omega t)]$, with frequencies and wavenumbers related by dispersion equations. For equation (1), the dispersion equation is:

$$\omega = uk \tag{6}$$

There is no diffusion because ω is real, and because the phase speed, $u = \omega/k$, is constant, there is no dispersion. The dispersion equations for computational schemes (3), (4), and (5), respectively, are:

$$\omega = \mathrm{i}\log\{[2 + \cos k\Delta - \gamma^2(1 - \cos k\Delta)/\tau - 3\mathrm{i}\gamma\sin k\Delta]/[2 + \cos k\Delta + 2\gamma^2(1 - \cos k\Delta)]\}/\tau \tag{7}$$

$$\omega = \mathrm{i}\log[1 - \gamma^2(1 - \cos k\Delta) - \mathrm{i}\gamma\sin k\Delta]/\tau \tag{8}$$

$$\omega = \mathrm{i}\log[1 - \gamma(1 - \cos k\Delta) - \mathrm{i}\gamma\sin k\Delta]/\tau \tag{9}$$

where $\gamma = u\tau/\Delta$ is the Courant number and $\mathrm{i} = (-1)^{\frac{1}{2}}$.

The real and imaginary parts of the frequency are shown in *Figures 1–3* as functions of the wave number. The fact that the real parts of equations (7)–(9) all deviate from equation (6) indicates that there are dispersive errors for all three computational schemes. By expanding, it can be readily seen that the imaginary part of the frequency for the upwind scheme is of order $(k\Delta)^2$, as for Fickian diffusion. On the other hand, because the imaginary parts of both equation (7) and (8) are of order $(k\Delta)^4$, the error minimizing scheme and the Leith–Lax–Wendroff schemes are only weakly diffusive, with much slower spreading than for Fickian diffusion.

Because the error functional is a quadratic function of the nodal values c_j^{n+1}, the effects of the constraints can easily be visualized. Since the error increases parabolically away from its minimum values in every direction, it follows that, when the minimum corresponds to negative values, the minimum for non-negative values must lie on the boundary of the positive sector. In other words, the non-negative minimum can be found from the negative absolute minimum by following the least steep path through the multidimensional space to the boundary of the positive sector. Roughly, this amounts to resetting the negative values to zero, but, in reality, all values are adjusted somewhat. Although the absolute minimum is conservative, the least steep path away from the minimum does not lie in the plane where conservation is satisfied. The non-negative minimum found by following the least steep path corresponds to discarding the unwanted negative values without also discarding an equal positive quantity. Therefore, to satisfy both non-negativity and conservation, it is necessary to follow the least steep path in the plane determined by the conservation constraint.

Conservation can be enforced through the use of a Lagrange multiplier. Positivity is an example of an inequality (Kuhn–Tucker) constraint and can be enforced through the use of slack variables,[10] but a simpler approach is taken here. When the solution to equation (1) is non-negative, the negative values resulting from equation (3) will be small so long as resolution is adequate. In the limit of infinite resolution, the negative values converge to zero. Thus, it is reasonable to assume that the non-negative minimum would have every negative value reset to zero (none should become positive). From this assumption, it is possible to derive finite difference equations to determine how the positive values should be readjusted.

The assumption that the path from the absolute minimum to the boundary of the positive sector intersects the portion of that sector where all the negative values are reset to zero is quite reasonable. It can be shown to be true for the case when there are less than five grid points, therefore, it might be possible to prove it for an arbitrary number of grid points. (For the special case of $\gamma^2 = 0.5$, it is easily shown to hold for any number of grid points.) In any case, any error due to this approximation should be of the same order as the increase in truncation error caused by replacing the absolute minimum with the non-negative minimum.

The first step is to solve equation (3) for the absolute minimum of the error functional. If there are no negative values, then the solution is both non-negative and conservative. If, however, there are any negative values, equation (3) must be replaced by a new set of equations: each variable that is negative or zero at the absolute minimum should be set equal to zero; for each variable that is positive at the absolute minimum, there is a corresponding equation in the form:

$$\frac{1}{\tau}\left[\frac{1}{6}(c_{j+1}^{n+1} + 4c_j^{n+1} + c_{j-1}^{n+1}) - \frac{1}{6}(c_{j+1}^n + 4c_j^n + c_{j-1}^n)\right] + \frac{u}{2\Delta}(c_{j+1}^n - c_{j-1}^n) = \frac{u^2\tau}{2}\left[\frac{2}{3\Delta^2}(c_{j+1}^{n+1} - 2c_j^{n+1} + c_{j-1}^{n+1}) + \frac{1}{3\Delta^2}(c_{j+1}^n - 2c_j^n + c_{j-1}^n)\right] - \frac{\lambda}{2\Delta} \tag{10}$$

and the Lagrange multiplier λ is determined by the conservation constraint:

$$\sum_{j=1}^{J} c_j^{n+1} = \sum_{j=1}^{J} c_j^0 \tag{11}$$

If there are any new negative values, then this second step must be repeated.

Equation (1) is identical to equation (3) except for the term involving the Lagrange multiplier. This term serves to reduce the excess in the positive values that results from discarding the unwanted negative values. When $\gamma = \pm(0.5)^{\frac{1}{2}}$, these equations are time-explicit and it is easy to see that each positive variable is reduced by the same amount. Whatever the Courant number, the compensation for discarding negative values is global rather than local. This should not be thought of as an 'instantaneous' redistribution of advected material

from great distances to fill in the negative values. Instead, it should be considered to be a redistribution of truncation error.

Suppose that there is a more or less constant, positive background level far from the region where negative vafues might develop in the absence of the positivity constraint. The effect of the Lagrange multiplier is then to slightly reduce the level of this background. The amount of this reduction is certainly less than the peak level of negative values discarded, because the compensation has been spread over the entire domain. The truncation error at the remote points is increased pointwise by less than the amount that it is reduced at points corresponding to negative values. However, because the absolute minimum is less than the non-negative minimum, the total truncation error is increased.

It is interesting to note that many computational schemes, the Leith–Lax–Wendroff scheme for example, can be considered to correspond to the minimum of some error functional. It is not necessary to derive the error function from some residual because it can be determined by integration. From this point of view, it is easy to modify such a scheme to incorporate the constraints of positivity and conservation simply by following the same procedure described above. In the next section computational results will be presented for such a non-negative Leith–Lax–Wendroff scheme.

Results

Results are presented for five computational schemes: the error minimizing scheme without correction for negative values; the error minimizing scheme with correction for negative values; the upwind scheme; the Leith–Lax–Wendroff scheme; and a Leith–Lax–Wendroff scheme modified to eliminate negative values.

In every case, a spatial lattice of 100 points was used, with initial conditions:

$$c_j^0 = \begin{cases} 1, j = 1, \ldots, 10 \\ 0, j = 11, \ldots, 100 \end{cases} \qquad (12)$$

As the short-wave components needed to form the corners of this initial distribution cannot be resolved by the spatial lattice, the simulations can be expected to be characterized by severe numerical dispersion. Graphical results depict the shape of the pulse at the half-way point as it propagates repeatedly from left to right across the periodic lattice. Results are shown for times $t_1 = 50\Delta/u$ and $t_2 = 550\Delta/u$, corresponding to the first and sixth transits of the grid respectively. The exact solution, namely propagation without change of shape, is also shown.

The curves in *Figure 4* correspond to the error minimization scheme without correction for negative values. Five cases are shown, corresponding to the same five values of the Courant number represented by the dispersion curves in *Figure 1*. The dispersion of the distribution into a wavetrain is evident at all Courant numbers. The computational results concur with expectations based on the dispersion curves: the shortest, most dispersed waves have been damped; the simulated speed of advection decreases with increasing Courant number; and the pulse is too poorly-resolved to be accurately simulated at $\gamma = 5$.

The curves in *Figure 5* correspond to the error minimization scheme with solutions constrained to be non-

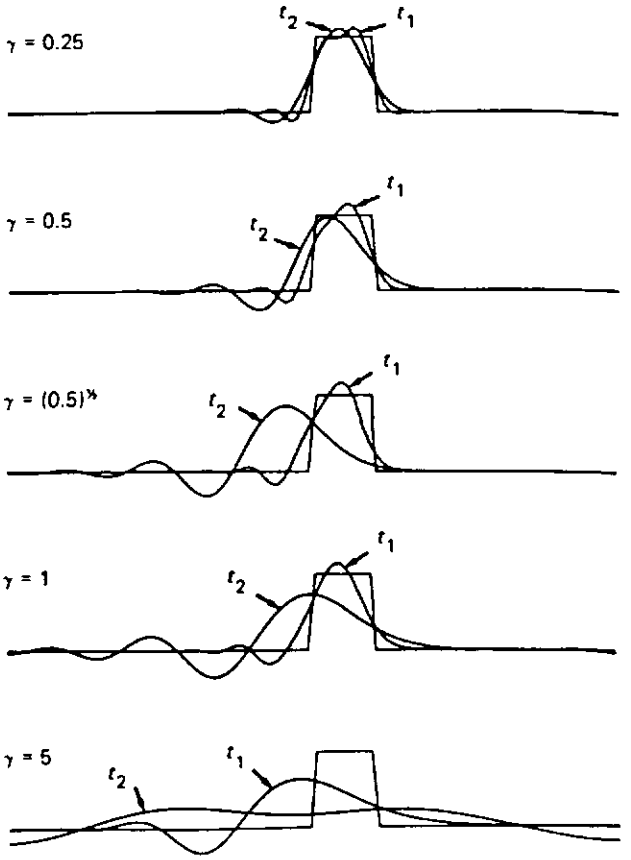

Figure 4 Computational results for propagation of initially trapzoidal pulse obtained using error-minimizing scheme for various Courant numbers. Curves shown for times $t_1 = 50\,\Delta/u$ and $t_2 = 550\,\Delta/u$, corresponding to midway point of first and sixth transits of pulse across periodic grid. Exact solution also shown for comparison

negative and conservative. It is interesting to note that, even though the error for each time-step is less when negative values are allowed, the results of many steps might be judged better in *Figure 5*. One possible reason for this is that correcting for negative values avoids systematic error that otherwise accumulates when stepping forward in time. Another is that, since the definition of error is based on the residual rather than directly on the solution, correcting for negative values might indeed give more accurate solutions. No effort has been made to investigate either of these possibilities. In any case, the non-negative results are very good.

Results for the upwind scheme are shown in *Figure 6*. The same cases are shown as in *Figures 4* and *5*, except that $\gamma = 5$ is omitted because of instability. For $\gamma = 1$, this scheme is exact, but for other values it is so highly diffusive that there is little resemblance between exact and computed solutions. Because diffusive errors dominate the dispersive errors, this scheme produces no negative values.

Although the upwind scheme is unstable for Courant numbers greater than unity, by enforcing conservation it is possible to use large time-steps and still not get exponentially-growing solutions. This does not mean, however, that the solutions will be reasonable. The solution will develop local spikes that become saturated after growing to the limit allowed by conservation.

Figure 7 presents results for the Leith–Lax–Wendroff scheme and *Figure 8* presents corresponding results for

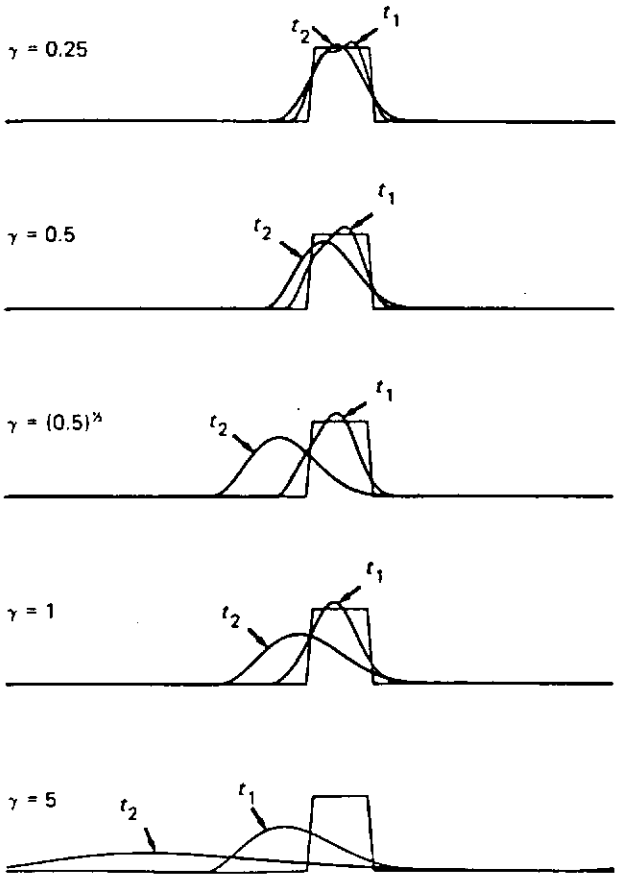

Figure 5 Computational results obtained using error-minimizing scheme and allowing only non-negative values for same cases as in *Figure 4*

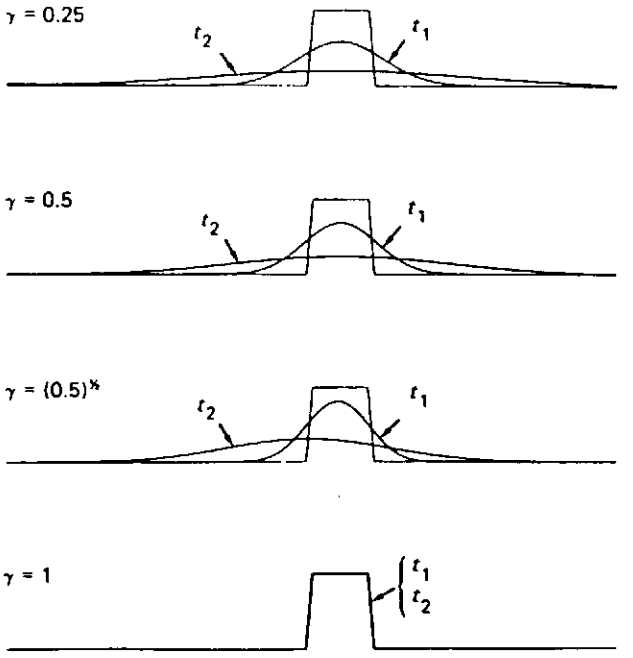

Figure 6 Computational results obtained using upwind scheme for same cases as in *Figures 4* and *5*, except no results shown for $\gamma = 5$

a Leith–Lax–Wendroff scheme modified to forbid negative values. The curves in *Figures 7* and *8* for $\gamma = (0.5)^i$ are identical to their counterparts in *Figures 4* and 5, since the schemes are identical for this Courant number. As for the upwind scheme, these results are exact for $\gamma = 1$; no results are shown for $\gamma = 5$ because of instability. The modified Leith–Lax–Wendroff scheme gives quite good results; correcting for negative values changes the solution very little where it agrees well with the exact solution and improves it where it disagrees most. Of the schemes discussed here, the non-negative

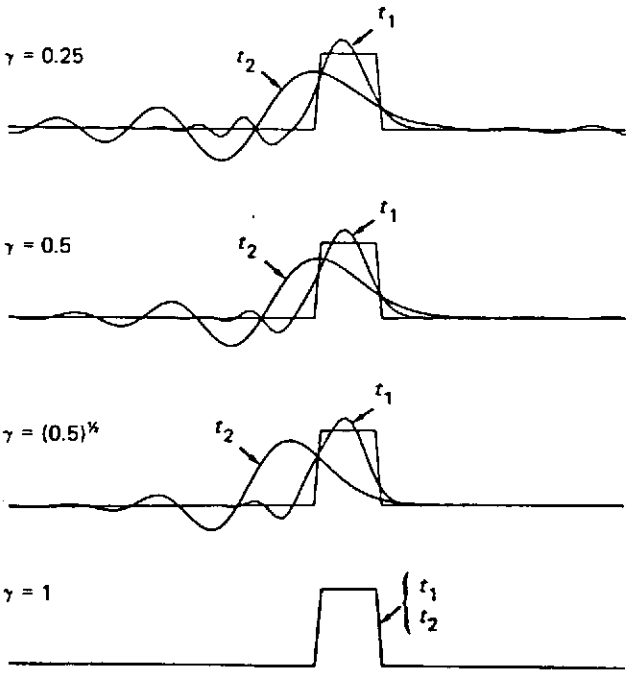

Figure 7 Computational results obtained using Leith–Lax–Wendroff scheme for same cases as in *Figure 6*

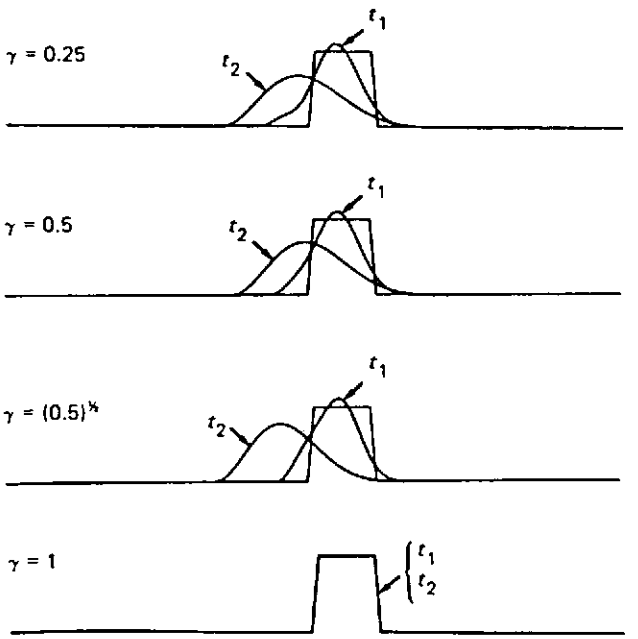

Figure 8 Computational results obtained using Leith–Lax–Wendroff scheme modified to allow only positive values for same cases as in *Figure 7*

Leith–Lax–Wendroff scheme provides the best accuracy for fixed computational expense. In addition, it should be competitive in comparison with other schemes.

Conclusions

This paper should be regarded as a comment on the nature of truncation error rather than as an advertisement for a particular computational scheme. All schemes necessarily exhibit truncation error; what makes one scheme less desirable than another is that its truncation errors take a form that is unacceptable for the intended application. The example discussed here is the advective transport of an intrinsically non-negative quantity, where the undesirable effects of truncation error are excessive diffusion and/or negative values. The approach has been to examine the consequences of minimizing a reasonable measure of truncation error under the constraints that the solution must be both non-negative and conservative.

This approach has produced three results. First, minimization of truncation error without regard to negative values or conservation has yielded a time-implicit finite difference scheme with attractive properties: It is quite accurate; it is conservative without imposing conservation as a constraint; it is only weakly diffusive; and the dissipation affects only those wave components that suffer from significant phase-speed errors. Because it is time-implicit, however, it is not likely to be cost-competitive for most applications.

The second result is a scheme that restricts the minimum truncation error to non-negative, conservative solutions. Put simply, this amounts to discarding the unwanted negative values produced by the unconstrained scheme and then reducing the positive values so that conservation is maintained. The details of how the positive values should be reduced are dictated by the error-minimization formalism. In practice, all that is required is an iterated solution to equations similar in form to the unconstrained equations, but with additional decay terms involving the Lagrange multiplier used to enforce conservation.

The third result is a prescription for enforcing non-negativity and conservation with any time-explicit advection scheme: Discard all negative values and compensate by reducing all positive values equally until conservation is recovered. This prescription is attractive because it is simple and easy to apply. In fact, it is the sort of thing that might have been done in the past for expediency, but without theoretical support.

The method of minimizing truncation error to derive finite difference equations is quite general and can be applied to a wide variety of problems. When applied to a set of coupled partial differential equations, the error density must be constructed from the residuals of each equation; an obvious choice is the sum of the squares of the residuals. However, because different physical quantities have different units, each term should have an appropriate dimensional coefficient. Examples of application of this method to the shallow water wave equations are discussed in references 11 and 12.

There is also no problem in applying this method to problems with two or three spatial dimensions. As the number of dimensions increases, however, and because the finite difference forms involve spatial and temporal averages, the number of terms in the resulting finite difference equations also increases. Another feature observed when this method is applied to partial-differential equations with many terms is the appearance of a multitude of finite difference terms for which there are no partial-differential counterparts. These additional terms, which vanish in the limit of infinitely fine grid spacing, are like the artificial-viscosity terms of the Lax–Wendroff method and their purpose is to ensure that truncation error is indeed minimized. The proliferation of terms, together with the fact that the equations are time-implicit, render error minimization *per se* unattractive for most applications.

The method for incorporating inequality and conservation constraints is also quite general and can also be used for a wide class of problems. For the method to be conceptually clear, it should be used in conjunction with truncation error density. As in the case of the Leith–Lax–Wendroff scheme, however, the error density does not have to be specified *a priori*, so the unattractive features (multiplicity of terms and time-implicit structure) can be avoided.

Acknowledgments

Thanks are due to the GKSS Forschungszentrum Geesthacht where the first author was a guest when this work was initiated and to the Alexander von Humboldt Stiftung for the Senior US Scientist Award, which made that visit possible.

References

1 Courant, R., Isaacson, E. and Rees, M. 'On the solution of nonlinear hyperbolic differential equations by finite differences', *Commun. Pure and Appl. Math.*, 1952, **5**, 243–255
2 Lax, P. D. and Wendroff, B. 'Systems of conservation laws', *Commun. Pure and Appl. Math.*, 1960, **13**, 217–237
3 Lax, P. D. and Wendroff, B. 'Difference schemes for hyperbolic equations with high order of accuracy', *Commun. Pure and Appl. Math.*, 1964, **17**, 381–398
4 Leith, C. E. 'Numerical simulation of the earth's atmosphere', *Methods Comput. Phys.*, 1965, **4**, 1–28
5 Patel, M. K. and Markatos, N. C. 'An evaluation of eight discretization schemes for two-dimensional convection–diffusion equations.' *Int. J. Numer. Methods Fluids*, 1986, **6**, 129–154
6 Patel, M. K., Markatos, N. C. and Cross, M. 'Method of reducing false-diffusion errors in convection–diffusion problems', *Appl. Math. Modelling*, 1985, **9** (4), 302–306
7 Boris, J. P. and Book, D. L. 'Flux-corrected transport—1: SHASTA, a fluid transport algorithm that works', *J. Comput. Phys.*, 1973, **11**, 38–69
8 Zalesak, S. T. 'Fully multidimensional flux-corrected transport', *J. Comput. Phys.*, 1979, **31**, 335–362
9 Zienkiewicz, O. C. 'The finite element method', McGraw-Hill, Maidenhead, 1977
10 Gue, R. L. and Thomas, M. E. 'Mathematical methods in operations research', Macmillan, New York, 1968
11 Petersen, P., Häuser, J., Thacker, W. C. and Eppel, D. 'An error minimizing scheme for nonlinear shallow water wave equations with moving boundaries, in 'Numerical methods for nonlinear problems', vol. 2, (ed. C. Taylor *et al.*), Pineridge Press, Swansea, 1984, pp. 826–836
12 Bahal, B., Thacker, W. C., Häuser, J. and Eppel, D. 'An error minimizing algorithm for the shallow-water wave equations', *Proc. Int. Conf. Accuracy Estimation and Adaptive Refinements in Finite Element Comput.*, Lisbon, Portugal, 1984, pp. 19–22