JOURNAL OF COMPUTATIONAL PHYSICS 21, 270–278 (1976)

Variational Design of Finite-Difference Schemes for Initial Value Problems with an Integral Invariant

YOSHI KAZU SASAKI

University of Oklahoma, Department of Meteorology, Norman, Oklahoma, 73069

Received September 2, 1975; revised December 29, 1975

A technique for deriving a finite-difference scheme to solve initial value partial-differential equations is presented. The set of partial-differential equations is assumed to possess one or more invariant integral quantities. In fluid dynamics, the integral of the total energy over the domain is frequently assumed invariant. This technique is based on the variational method and constrains the finite-difference scheme to satisfy the conservation law(s). The technique is discussed by considering, as an example, a set of linear, shallow-water equations on a rotating plane, and extending it to a nonlinear case.

1. INTRODUCTION

The success of the numerical simulation of fluid dynamics and geophysical fluid dynamics, depends upon the proper design of a numerical finite-difference scheme for the integration of the governing differential equations. Beyond the problems of consistency, convergence, and stability of solutions (Richtmyer [10]) is another problem that arises when the governing differential equations have an auxiliary condition(s) which is expressed as an integral invariant(s). The integral invariant(s) in fluid and geophysical fluid dynamics is often represented by a conservation law(s). A numerically integrated solution of finite-difference equations may be obtained without considering the auxiliary conservation law(s), but the solution may violate the conservation condition (Lax and Wendroff [7]).

Arakawa [1] pointed out the importance of developing a scheme to conserve quadratic quantities such as kinetic energy in atmospheric prediction models. In the words of Arakawa [1], "When quadratic quantities are conserved in a finite-difference scheme, nonlinear computational instability cannot occur. This follows from the fact if the square of a quantity is conserved with time when summed up over all the grid points in the domain, the quantity itself will be bounded, at every individual grid point, throughout the entire period of integration." The computational instability is caused by the growth of error due to aliasing (Phillips, [9]). It

should be noted, however, that while quadratic conservation schemes are able to put a bound on the amplitude of noise waves, they are not able to control their phase. In spite of this limitation, considerable effort has been made to satisfy the conservation principles when designing finite-difference approximations of the governing differential equations. Arakawa [1] has developed a spatial finite-difference scheme for the two-dimensional vorticity equation that retains the integral invariants of the continuous system. The integral invariants are kinetic energy, squared vorticity (enstrophy) and vorticity. Lilly [8] developed a finite-difference form of the nonlinear terms for two-dimensional barotropic flow which retains spatial momentum and total energy conservation. Smagorinsky, Manabe, and Holloway [17] used it in their general circulation model. A generalized version of the Arakawa scheme for a shallow-water fluid is given by Grammeltvedt [3]. Shuman [15] and Shuman and Vanderman [16] developed a scheme that for practical purposes conserves total energy. Energy conserving schemes have also been developed individually by Bryan [2], Grimmer and Shaw [4] and Kurihara and Holloway [6] for curvilinear grids. Grammeltvedt [3] presented a detailed discussion and very useful comparison among the schemes used in numerical weather prediction. A comprehensive review of fluid dynamical schemes was given by Roache [11].

All of these conservation schemes are in rather complicated finite-difference forms which are difficult to extend to some fluid dynamically sophisticated problems of interest. For instance, complication arises in trying to impose the conservation requirements on implicit or semi-implicit differencing schemes. None of the existing so-called "conservation schemes" strictly satisfies the required conservation law(s) for long time integrations, as demonstrated by Grammeltvedt [3]. It is highly desirable, therefore, that a systematic approach for designing simple conservative numerical schemes be developed, explicitly enforcing the required conservation relationships. The following approach is based on the variational method (Sasaki [12, 13], Stephens [18]).

2. Variational Design I (with a Linear Model)

As an example, we consider the motion of shallow-water gravity waves on a rotating plane, under the condition that the total energy is conserved. First, for simplicity, the linearized governing equations (Haltiner [5]) are considered:

$$(\partial u/\partial t) - fv = 0, \tag{1}$$

$$(\partial v/\partial t) + fu + g(\partial h/\partial y) = 0, \tag{2}$$

and

$$(\partial h/\partial t) + H(\partial v/\partial y) = 0. \tag{3}$$

In these equations f, g, and H are the Coriolis parameter, the gravity acceleration, and the mean depth of the shallow fluid, respectively; t is time; x and y are Cartesian space coordinates; u and v are, respectively, the x and y components of water particle velocity. The elevation of the free water surface, measured from the mean height, is designated by h. The first and second equations are the linearized perturbation equations of motion, which are derived by assuming that the mean state is a motionless, homogeneous layer of water and the perturbations u, v, and h are uniform in the x direction.

This shallow water system conserves the total energy (the integral along y of the perturbation kinetic energy $H(u^2 + v^2)/2$ and available potential energy $gh^2/2$) when no energy accumulation occurs due to net total energy fluxes through the lateral boundaries $y = y_1$ and y_2. In the domain bounded by the lateral boundary, the total energy conservation is written

$$\int_{y_1}^{y_2} [(H/2)(u^2 + v^2) + (g/2)\, h^2]\, dy = \langle T \rangle, \tag{4}$$

where $\langle T \rangle$ is a constant. This invariant law is the most important equation which will be used in designing a new numerical scheme.

The set of Eqs. (1)–(3) will be solved as an initial value problem by specifying initial conditions for u, v, and h. Let us place a lattice on the time and space (y) domain. The time and space intervals of the lattice are designated by Δt and Δy respectively. A point of the lattice is represented by n and j which denote the nth time level and the jth space location, respectively, and u, v, and h are assigned to all points of the lattice. Then we choose a finite difference analogue of (1)–(3).

The variables at the $(n + 1)$th time level are determined by this set of finite-difference equations when the variables at the nth time level are all known. These predicted variables are denoted by $\tilde{u}$, $\tilde{v}$, and $\tilde{h}$. They are determined uniquely without considering the requirement of total energy conservation (4) which is satisfied by the true solution of the differential equations (1)–(3). Since the $\tilde{u}$, $\tilde{v}$ and $\tilde{h}$ may contain truncation error, we should be allowed to adjust them slightly to satisfy the required conservation law. The total energy (TE) computed from $\tilde{u}$, $\tilde{v}$, and $\tilde{h}$,

$$TE = \sum \left[H \frac{(\tilde{u}^2 + \tilde{v}^2)}{2} + \frac{g\tilde{h}^2}{2} \right] \tag{5}$$

does not satisfy the conservation law (4),

$$TE \neq T^0 \tag{6}$$

where T^0 is TE at $t = 0$ and $\sum$ represents a summation over all gridpoints along y. It is now conceivable to attempt to modify $\tilde{u}$, $\tilde{v}$, and $\tilde{h}$ to satisfy the energy conservation at each time step.

In this variational design of a numerical method, three basic hypotheses are used. At this point we will cover the first two. The first hypothesis follows.

Conservation laws valid for the true solution of the differential equations should also hold for the finite-difference solution.

Using this principle, (4) should be written in the following form:

$$\sum [(H/2)(u^2 + v^2) + (g/2)\, h^2] = T^0, \tag{7}$$

where T^0, the total energy, may be determined from initial values alone. The second hypothesis relates the solution of (7) to the forecast values.

The solution (u, v, and h) is also a stationary value that minimizes a weighted sum of the variances of ($u - \tilde{u}$), ($v - \tilde{v}$) and ($h - \tilde{h}$), integrated over the entire domain.

Based on these two hypothesis, a variational formulation of the problem is now easily made.

The functional

$$J = \sum [\tilde{\alpha}(u - \tilde{u})^2 + \tilde{\alpha}(v - \tilde{v})^2 + \tilde{\beta}(h - \tilde{h})^2] + \lambda_E \left\{ \sum [(H/2)(u^2 + v^2) + (g/2)\, h^2] - T^0 \right\} \tag{8}$$

will have a stationary value if its first variation equals zero,

$$\delta J = 0, \tag{9}$$

where δ is the variational operator. In (8), the summation $\sum$ is taken over all j points, $\tilde{\alpha}$ and $\tilde{\beta}$ are weights which will be determined later, and λ_E is the Lagrange multiplier which is constant with respect to space but possibly varies in time. Taking the first variation of (8) with respect to u, v, h and λ_E,

$$\delta J = \sum [2\tilde{\alpha}(u - \tilde{u})\, \delta u + 2\tilde{\alpha}(v - \tilde{v})\, \delta v + 2\tilde{\beta}(h - \tilde{h})\, \delta h + \lambda_E H(u\, \delta u + v\, \delta v) + \lambda_E g h\, \delta h] + \delta\lambda_E \left\{ \sum [(H/2)(u^2 + v^2) + (g/2)\, h^2] - T^0 \right\}.$$

After rearranging the equation,

$$\delta J = \sum [(2\tilde{\alpha}(u - \tilde{u}) + \lambda_E H u)\, \delta u + (2\tilde{\alpha}(v - \tilde{v}) + \lambda_E H v)\, \delta v + (2\beta(h - h) + \lambda_E g h)\, \delta h] + \left\{ \sum [(H/2)(u^2 + v^2) + (g/2)h^2] - T^0 \right\} \delta\lambda_E . \tag{10}$$

Since the variations δu, δv, δh and $\delta\lambda_E$ are arbitrary values (nonzero), the coeffi-

cients of each variation must vanish individually in order to satisfy the stationarity condition. This leads to the so-called Euler-Lagrange equations:

$$u = \frac{2\tilde{\alpha}}{2\tilde{\alpha} + \lambda_E H}\,\tilde{u}, \tag{11}$$

$$v = \frac{2\tilde{\alpha}}{2\tilde{\alpha} + \lambda_E H}\,\tilde{v}, \tag{12}$$

$$h = \frac{2\tilde{\beta}}{2\tilde{\beta} + \lambda_E g}\,\tilde{h}, \tag{13}$$

and

$$\sum [(H/2)(u^2 + v^2) + (g/2)\,h^2] = T^0. \tag{14}$$

Note that the last equation (14) is simply the total energy constraint, Eq. (7).

Now the third hypothesis is introduced, which will concern the determination of the weights $\tilde{\alpha}$ and $\tilde{\beta}$. One of the weights can be taken to be unity, say $\tilde{\alpha} = 1$. The only the ratio of $\tilde{\beta}$ to $\tilde{\alpha}$, called a relative weight, need be determined. The third hypothesis is therefore concerned with the determination of the relative weight:

The relative weight is chosen to make the fractional adjustment of variables proportional to the fractional magnitude of the truncation errors in the predicted variables.

The fractional adjustment of u and v is $2\tilde{\alpha}/(2\tilde{\alpha} + \lambda_E H)$ and that for h is $2\tilde{\beta}/(2\tilde{\beta} + \lambda_E g)$. If the u, v, and h trunctation errors have the order of magnitude $O(\Delta t^2, \Delta y^2)$, the hypothesis leads to

$$\frac{2\tilde{\alpha}}{2\tilde{\alpha} + \lambda_E H} = \frac{2\tilde{\beta}}{2\tilde{\beta} + \lambda_E g} \equiv X, \tag{15}$$

where X is the fractional adjustment rate. From (15), we can easily solve for the relative weight,

$$\tilde{\beta}/\tilde{\alpha} = g/H. \tag{16}$$

If, for instance we chose $\tilde{\alpha} = 1$, $\tilde{\beta}$ becomes g/H and the weights $\tilde{\alpha}$ and $\tilde{\beta}$ are specified.

We now have four unknowns, namely, u, v, h, and λ_E, and four Euler–Lagrange equations (11)–(14). Through Eq. (15), the variable λ_E may be replaced by X. The unknown X is solved from the equation which will be derived by substituting u, v, and h from (11)–(13) into (14), giving

$$X^2 \sum [(H/2)(\tilde{u}^2 + \tilde{v}^2) + (g/2)(\tilde{h}^2)] = T^0,$$

or

$$X = \left\{T^0 \Big/ \sum [(H/2)(\tilde{u}^2 + \tilde{v}^2) + (g/2)(\tilde{h}^2)]\right\}^{1/2}. \tag{17}$$

Substitution of this X into (11)–(13) yields the solutions u, v and h, respectively. The Lagrange multiplier λ_E is given by (15) as function of X,

$$\begin{aligned} \lambda_E &= 2\tilde{\alpha}[(1/X) - 1]/H \\ &= 2\tilde{\beta}[(1/X) - 1]/g. \end{aligned} \tag{18}$$

3. Variational Design II (with a Nonlinear Model)

The technique described above for a linear equation system may be extended to nonlinear systems. A simple example will be shown in this chapter.

We consider a nonlinear form of (1–3),

$$\frac{\partial(hu)}{\partial t} + \frac{\partial}{\partial y}(v\,hu) - f\,hv = 0, \tag{19}$$

$$\frac{\partial(hv)}{\partial t} + \frac{\partial}{\partial y}(v\,hv) + f\,hu + g\,\frac{\partial}{\partial y}\left(\frac{h^2}{2}\right) = 0, \tag{20}$$

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial y}(hv) = 0, \tag{21}$$

where h is the depth of the fluid, which was the height anomaly in the linear system, but is now the total depth in the nonlinear system.

The energy conservation law is written as

$$\int_{y_1}^{y_2} [(h/2)(u^2 + v^2) + (g/2)\,h^2]\,dy = \langle T \rangle, \tag{22}$$

where $\langle T \rangle$ is a constant. The conservation may be written in a finite difference analog as:

$$TE \equiv \sum [(h/2)(u^2 + v^2) + (g/2)\,h^2] = T^0, \tag{23}$$

where T^0 is the value of TE at $t = 0$. Let $\tilde{h}$, $\tilde{u}$, and $\tilde{v}$ be the values predicted for the $(n + 1)$th time level by using a set of finite difference equations corresponding to (19)–(21).

We now formulate the variational problem as we did for the linear system. The functional J becomes

$$\begin{aligned} J = \sum [\tilde{\alpha}(u - \tilde{u})^2 + \tilde{\alpha}(v - \tilde{v})^2 + \tilde{\beta}(h - \tilde{h})^2] \\ + \lambda_E \left\{\sum [(h/2)(u^2 + v^2) + (g/2)\,h^2] - T^0\right\}. \end{aligned} \tag{24}$$

The stationary value of the functional results from setting its first variation to zero. The resulting Euler–Lagrange equations are

$$2\tilde{\alpha}(u - \tilde{u}) + \lambda_E h u = 0, \tag{25}$$

$$2\tilde{\alpha}(v - \tilde{v}) + \lambda_E h v = 0, \tag{26}$$

$$2\tilde{\beta}(h - \tilde{h}) + \lambda_E[(u^2 + v^2)/2] + \lambda_E g h = 0, \tag{27}$$

and

$$\sum [(h/2)(u^2 + v^2) + (g/2)\, h^2] - T^0 = 0. \tag{28}$$

The numerical solutions u, v, h, and λ_E may be obtained by an iterative technique. The author tested a semilinear iterative technique. First, assuming $H \approx h$, the solutions u, v, and h are approximated as follows:

$$u \doteq [2\tilde{\alpha}/(2\tilde{\alpha} + \lambda_E H)]\, \tilde{u}, \tag{29}$$

$$v \doteq [2\tilde{\alpha}/(2\tilde{\alpha} + \lambda_E H)]\, \tilde{v}, \tag{30}$$

$$h \doteq [2\tilde{\beta}/(2\tilde{\beta} + \lambda_E g)]\, \tilde{h}. \tag{31}$$

Note that these expressions have the same form as those of the linear system. Additional simplification is obtained by limiting the magnitude of λ_E to be sufficiently small that the equation for λ_E obtained by substituting u, v and h from (25)–(27) into (28) is linear. The magnitude of λ_E depends on the variances $(u - \tilde{u})^2$, $(v - \tilde{v})^2$, and $(h - \tilde{h})^2$. If the finite-difference scheme is fairly accurate, large variances should not occur in only one time step.

If the difference between T^0 and the total energy TE computed from the forecasted variables was less than an arbitrary limit (10^{-7} times T^0 was used), no adjustment was made. For typical differences, the error was reduced below the limit in one or two iterations.

Another integral invariant appropriate to equations (19)–(21) is the total mass.

$$\int_{y_1}^{y_2} h\, dy = \text{constant} \tag{32}$$

or in a finite difference analog,

$$\sum h = h^0, \tag{33}$$

where h^0 is the value of $\sum h$ at $t = 0$. This constraint could be enforced by adding

$\lambda_h(\sum h - h^0)$ to the functional J, where λ_h is another Lagrange multiplier. However, a simpler approach has been found successful for this experiment. The mass conservation constraint was treated separately from the total energy constraint. The depths forecasted by the finite-difference equations were first adjusted to achieve mass conservation as follows:

$$h = \tilde{h} - \left(\sum \tilde{h} - h^0\right) \Big/ j_m , \tag{34}$$

where $j = 1, 2, \ldots, j_m$. Then using these corrected values for $\tilde{h}$ in (24), the total energy constraint was applied as described above. The entire process is summarized below:

1. Computation of $\tilde{h}$, $\tilde{u}$, $\tilde{v}$ by a set of prognostic finite-difference equations.
2. Adjustment of $\tilde{h}$ to ensure total mass conservation.
3. Solution of the Euler–Lagrange equations to enforce the total energy conservation.

Steps 1, 2, and 3 were taken for each time step in the experiments. Then the total mass was rechecked. If a significant error occurred, the mass adjustment was repeated and the total energy rechecked. Normally, no more than two repetitions of this procedure were necessary. Satisfactory results could probably be obtained by applying the adjustments less frequently, although this idea has not been tested.

4. Remarks

Numerical tests of the variational scheme with the linear system (1)–(3) showed quite satisfactory results compared with the corresponding analytical solutions in terms of numerical accuracy and computer speed. One of the tests was made by using the so-called forward scheme to compute $\tilde{u}$, $\tilde{v}$, and $\tilde{h}$. The forward scheme employs centered space differencing and forward time differencing for the space and time derivatives, respectively. The scheme is known to be computationally unstable, but the total energy constraint puts a bound on the amplitude of $\tilde{u}$, $\tilde{v}$, and $\tilde{h}$, prohibiting the instability (Sasaki, [14]). The unstable scheme, however, amplifies the shorter waves more than the longer waves. In another experiment, the spurious short wave amplification was to a certain extent avoided by applying simple arithmetic smoothing to $\tilde{h}$, $\tilde{u}$, and $\tilde{v}$ before enforcing the total energy constraint. Some nonlinear experiments were conducted in which the governing Eqs. (19)–(21) were integrated under the constraints of total energy (22) and total mass (32). The numerical integration followed the procedures described in Section 3. Although no analytical solution is available to compare with the numerical solution, the numerically integrated results are quite satisfactory.

ACKNOWLEDGMENTS

This study was supported by the National Science Foundation under NSF Grant GA-30976. This manuscript was prepared while the author was at the University of Oklahoma, and revised while on sabbatical leave of absence from the University and working at the Environmental Prediction Research Facility, Naval Postgraduate School, Monterey, California. The author would like to express his appreciation to the staff of the above institutions for the assistance given in the preparation of this manuscript. Special thanks are extended to Professor George Haltiner (Department of Meteorology, Naval Postgraduate School), Mr. Edward Barker (Environmental Preduction Research Facility, Naval Postgraduate School) and Mr. Tom Baxter (University of Oklahoma) for their critical reading of the manuscript.

The author acknowledges personal communication from Professor Eugene Isaacson of the Courant Institute of Mathematical Science, New York University. Doctors Alvin Bayliss and Eugene Isaacson arrived at the idea of modifying the predicted values independently by using the constraints of conserving total mass and total energy, similar to the one described in this manuscript. The difference is that their approach linearized the constraints about the predicted values and the author's approach uses the Lagrange multiplier by the variational method. The author would like to express his thanks to Professor Isaacson for informing him of these interesting and independent developments.

REFERENCES

1. A. ARAKAWA, *J. Comput. Phys.* **1** (1966), 119.
2. K. BRYAN, *Mon. Wea. Rev.* **94** (1966), 39.
3. A. GRAMMELTVEDT, *Mon. Wea. Rev.* **97** (1969), 384.
4. M. GRIMMER AND D. B. SHAW, *Quart. J. Roy. Meteor. Soc.* **93** (1969), 337.
5. G. J. HALTINER, "Numerical Weather Prediction," Wiley, New York, 1971.
6. Y. KURIHARA AND J. L. HOLLOWAY, JR., *Mon. Wea. Rev.* **95** (1967), 509.
7. P. D. LAX AND B. WENDROFF, *Comm. Pure Appl. Math.* **13** (1960), 217.
8. D. K. LILLY, *Mon. Wea. Rev.* **93** (1965), 11.
9. N. A. PHILLIPS, "The Atmosphere and Sea in Motion," Rockefeller Institute Press, New York, 1959.
10. R. D. RICHTMYER, "Difference Methods for Initial-Value Problems," Interscience, New York, 1957.
11. P. J. ROACHE, "Computational Fluid Dynamics," Hermosa Publishers, Albuquerque, New Mexico, 1972.
12. Y. K. SASAKI, *J. Meteor. Soc. Japan* **33** (1955), 262.
13. Y. K. SASAKI, Proceedings of the WMO/IUGG Symposium on Numerical Weather Prediction, Tokyo, Japan, Japan Meteorological Agency, pp. VII-25, 1969.
14. Y. K. SASAKI, "Variational Design of Finite Difference Scheme for Initial Value Problem of Conservative System," Univ. Oklahoma, Norman, Oklahoma, 1975.
15. F. G. SHUMAN, Proceedings of the International Symposium on Numerical Weather Prediction in Tokyo, Meteorological Society of Japan, p. 85, 1962.
16. F. G. SHUMAN AND L. W. VANDERMAN, *Mon. Wea. Rev.* **94** (1966), 329.
17. J. SMAGORINSKY, S. MANAGE, AND J. L. HOLLOWAY JR., *Mon. Wea. Rev.* **93** (1965), 727.
18. J. J. STEPHENS, Ph.D. dissertation, Univ. Texas, Austin, Texas, 1965.