Astron. Astrophys. 108, 76–84 (1982)

ASTRONOMY AND ASTROPHYSICS

A Comparative Study of Computational Methods in Cosmic Gas Dynamics

G. D. van Albada[1]*, B. van Leer[2]**, and W. W. Roberts, Jr.[3]***

[1] The National Radio Astronomy Observatory, Charlottesville, VA, USA and Kapteyn Laboratory, Groningen, The Netherlands
[2] Institute for Computer Applications in Science and Engineering NASA Langley Research Center, Hampton, VA, and Leiden State University, Leiden, The Netherlands
[3] Institute for Computer Applications in Science and Engineering NASA Langley Research Center, Hampton, VA, and The University of Virginia, Charlottesville, VA

Received August 25, accepted December 11, 1981

Summary. In search of reliable computational methods for cosmic flow problems, we apply several commonly used algorithms and one new algorithm, to a representative problem in galactic gas dynamics. A careful choice of the algorithm used in a calculation is found to be of the utmost importance in obtaining reliable results. Two methods most commonly employed in astronomy (the Beam scheme and FCT methods) prove to be highly unsuitable for our test problem. The penalty in programming effort and computer time per grid point required for the best second-order accurate codes tested is more than offset by the improvement in accuracy obtained and the possibility to reduce the number of points in a grid.

Key words: hydrodynamics – hydromagnetics – interstellar matter – spiral galaxies

I. Introduction

Many theoretical investigations of astrophysical fluid flows require extensive numerical calculations. The techniques of computational fluid dynamics have been employed in virtually every area from stellar structure to cosmology. The use of numerical calculations in astrophysics can only increase as the capacity to perform substantial calculations becomes more widespread. Indeed, the computation of two-dimensional flows may be considered almost routine. Thus, increasing numbers of astronomers will be asking that basic question of scientific computing: What reliable, accurate, efficient, and easy-to-program method should be used for this calculation?

The average astronomer with only a casual knowledge of numerical methods is likely to confine his choices to methods that have been used before on problems in his particular field of research and to methods described in standard textbooks on numerical analysis. In the former case he is likely to be exposed to a one-sided defense of a single method. In the latter case he is unlikely to encounter any comparison between methods on a physical problem that embodies some of the special difficulties posed by astrophysical flows.

The objective of this article is to supply a comparison of a variety of numerical methods on a representative astrophysical flow problem. The methods considered include some that are widely used in astronomy, some that are widely used in other fields such as aerodynamics, and one that is rather new. All the methods are explicit and are especially suited for transonic and supersonic flows. This comparison aims to acquaint astronomers with the virtues and failings of typical numerical methods.

II. The Problem

A simple, one-dimensional model of the gas flow in a spiral galaxy will serve as the test problem. Numerical calculations of the nonlinear response of the gas to a mild spiral structure in the more massive stellar component of a disk galaxy have played a major role in establishing the density-wave theory of Lin and Shu (1964, 1966) as a viable explanation for the coherent, large-scale spiral patterns observed in many galaxies. [Recent reviews of this topic have been given by Toomre (1977) and by Lin and Lau (1979).] Although the dynamics of spiral galaxies are dominated by the far more massive (and much hotter) stellar component, their appearance is largely a result of the fierce response of the (cold) gas to the stellar gravitational field. The ease with which a mild stellar spiral structure can induce shock waves in the gas, and its implications for the observed features of spiral structure, was demonstrated by Roberts (1969) using one-dimensional, steady-state gas equations which included a forcing term due to the spiral field of the stars. The actual evolution of such flows was studied by Woodward (1975) using a simplified time-dependent version of Roberts' equations. We use Woodward's equations and also a set of his parameter values.

The nonlinear response of the gas to an imposed spiral gravitational field has several of the distinguishing characteristics of astrophysical flows. A major role is played by source terms, strong shocks are apt to develop, and rotational effects are significant. The presence of source terms can lead to unexpected behavior in numerical methods which are usually analyzed and tested in the absence of such terms. Strong shocks demand reliability; the proper treatment of angular momentum requires accuracy. Many methods cope with the shock only at the price of

Send offprint requests to: G. D. van Albada (Groningen)

* The National Radio Astronomy Observatory is operated by Associated Universities, Inc., under contract with the National Science Foundation
** This work was supported under NASA Contract No. NAS1-14472 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665
*** This work was supported in part by the National Science Foundation under Grant AST-7909935; partial support was also received under NASA Contract No. NAS1-15810 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA

artificially redistributing the angular momentum. This can be a serious problem if one is investigating the dynamics of the gas, since the physical system can be very sensitive to redistribution of angular momentum.

The equations in an inertial frame for an isothermal gas are

$$\frac{\partial \varrho}{\partial t}+\nabla\cdot(\varrho \boldsymbol{q})=0, \tag{1}$$

$$\frac{\partial \boldsymbol{q}}{\partial t}+\boldsymbol{q}\cdot\nabla\boldsymbol{q}=-\frac{c^2}{\varrho}\nabla\varrho-\nabla\Phi, \tag{2}$$

where ϱ is the density, $\boldsymbol{q}$ is the velocity, c is the (constant) sound speed and Φ is the gravitational potential. The isothermal assumption is used since interstellar gas cools by radiative processes on a much shorter time-scale than that of any dynamical processes. It is customary to use for c an equivalent dispersive speed which is partly thermal, partly due to turbulence, and partly due to cosmic ray particles.

In the absence of the spiral forcing the gas flow is circular with angular velocity $\Omega(r)$ at radius r. A steady spiral field with small pitch angle α is assumed to rotate rigidly with pattern speed Ω_p. A convenient coordinate system is one which rotates at this speed and is aligned with the equipotential contours of the spiral. The coordinates parallel and perpendicular to the equipotential contours are denoted by ξ and η, respectively. The velocity components in this frame are written as

$$\begin{aligned} v&=q_\xi,\\ u&=q_\eta. \end{aligned} \tag{3}$$

If we assume that the spiral has a pitch angle $\alpha\ll 1$, the equilibrium velocities are approximately

$$\begin{aligned} v_0&=r(\Omega-\Omega_p),\\ u_0&=\alpha r(\Omega-\Omega_p). \end{aligned} \tag{4}$$

In this approximation derivatives with respect to η (normal to the spiral arms) are retained, but derivatives with respect to ξ (along the spiral arms) are discarded. For a two-armed spiral the resulting equations can be written as the system of conservation laws

$$\frac{\partial U}{\partial t}+\frac{\partial F}{\partial \eta}=H, \tag{5}$$

where the vector of conserved quantities is

$$U=\begin{pmatrix}\varrho\\ \varrho u\\ \varrho v\end{pmatrix}, \tag{6}$$

the vector of fluxes is

$$F=\begin{pmatrix}\varrho u\\ \varrho(u^2+c^2)\\ \varrho uv\end{pmatrix}, \tag{7}$$

and the vector of source terms is

$$H=\begin{pmatrix}0\\ 2\Omega(v-v_0)\varrho+\dfrac{2}{\alpha r}\varrho A\sin\hat{\eta}\\ -\dfrac{\kappa^2}{2\Omega}(u-u_0)\varrho\end{pmatrix}. \tag{8}$$

The spiral phase $\hat{\eta}$ is defined by

$$\hat{\eta}=\frac{2\eta}{\alpha r}, \tag{9}$$

and the epicyclic frequency κ by

$$\kappa^2=\frac{2\Omega}{r}\frac{d}{dr}(r^2\Omega). \tag{10}$$

In this approximation the flow is periodic; in terms of the spiral phase the periodicity condition reads

$$U(\hat{\eta},t)=U(\hat{\eta}+2\pi,t). \tag{11}$$

The driving term $(2/\alpha r)\varrho A\sin\hat{\eta}$ arises from the assumed gravitational field of the stellar component. For the test problem we adopt the parameters thought to be appropriate for the neighborhood of the Sun in our own galaxy: $\Omega=25$ km s^{-1}/kpc, $\kappa=31.3$ km s^{-1}/kpc, $\Omega_p=13.5$ km s^{-1}/kpc, $c=8.56$ km s^{-1}, $r=10$ kpc, and $\alpha=\sin(6\overset{\circ}{.}7)\approx 0.11667$. For the amplitude A we choose $A=72.92$ (km s^{-1})^2, which makes the amplitude of the spiral force 2 % of the equilibrium force $r\Omega^2$.

In the steady state the Eqs. (5) become

$$\frac{du}{d\eta}=\frac{u}{u^2-c^2}\left[2\Omega(v-v_0)+\frac{2A}{\alpha r}\sin\frac{2\eta}{\alpha r}\right], \tag{12}$$

$$\frac{dv}{d\eta}=-\frac{\kappa^2}{2\Omega}\frac{u-u_0}{u}. \tag{13}$$

The particular case studied here becomes supersonic with a sonic point at spiral phase $\hat{\eta}=155\overset{\circ}{.}53$ and a shock at $\hat{\eta}=131\overset{\circ}{.}68$. A procedure for solving the steady state Eqs. (12) and (13) plus the periodicity condition (11) is described in Roberts (1969); see also Shu et al. (1973). The noteworthy features of this flow are the rapid decompression after the shock and the secondary structure near a spiral phase of 270° which is caused by resonance effects. The time-dependent version of this problem challenges a numerical method to cope with the shock, while also resolving the remaining structure of the flow.

III. The Methods

A typical numerical method for the system (5) divides the spatial region $(0,\pi\alpha r)$ into N zones centered at the grid points $=(i-\frac{1}{2})\Delta\eta$, where $i=1,2,\ldots,N$ and $\Delta\eta=\pi\alpha r/N$ (or $\Delta\hat{\eta}=2\pi/N$), and advances the approximate solution from time t_n to time t_{n+1} (where $t_{n+1}=t_n+\Delta t$) by means of a discretized version of the partial differential equations. The approximate value of U at the point (η_i,t_n) is denoted by U_i^n; F_i^n and H_i^n are defined as $F(U_i^n)$ and $H(U_i^n,\eta_i)$ or, as explained below, $H(U_i^n,\eta_i,\Delta t)$. In all methods discussed here, U_i^n actually approximates the average value of the solution over zone i. A subscript $i+\frac{1}{2}$ denotes an interpolated numerical value at the zone boundary $\eta=i\Delta\eta$, or a finite difference across this boundary; a superscript $n+\frac{1}{2}$ denotes an approximate value at $t_{n+1/2}=(n+\frac{1}{2})\Delta t$.

Our sample of numerical methods includes

a) The beam scheme (B) [see Sanders and Prendergast (1974)].

b) Godunov's (1959) method (G).

c) Second-order flux-splitting method (FS2) [see Van Leer (1981a)].

d) MacCormack's (1969) method (MC2).

e) Flux Corrected Transport (FCT) methods of Boris and Book (1973).

Before discussing these methods at length, some general remarks are due. The methods (a) and (b) are first-order accurate, that is, these approximate Eq. (5) with an error $O(\Delta x)$. The remaining methods are of the second order of accuracy. All first-order methods and also method (c) are based on upwind differenc-

ing. This means that, in approximating $\partial F/\partial\eta$ in (5), a distinction is made between contributions from wave motion or material motion in the positive direction and in the negative direction [see the review by Harten et al. (1981)]. The methods (d) and (e) are based on central differencing, in which the above distinction is not made.

The schemes can be written in the form

$$\frac{U_i^{n+1}-U_i^n}{\Delta t}+\frac{\Phi^v_{i+1/2}-\Phi^v_{i-1/2}}{\Delta\eta}=H_i^{n+1/2}, \tag{14}$$

with $v=n$ for the first-order methods and $v=n+\frac{1}{2}$ for the higher order methods. The latter all are two-step algorithms in which time centering is achieved using the first-order accurate results of the first step at $t_{n+1/2}$ or t_{n+1}.

The numerical flux vector

$$\Phi^v_{i+1/2}\equiv\Phi(U^n_{i-k+1},\ \ldots,\ U^n_{i+k}), \tag{15}$$

is a function of $2k$ arguments (initial values); for the time-step in question it is a representative value of F at $\eta_{i+1/2}$ according to some model of the interaction of the gas cells. It is the particular choice of Φ that distinguishes one scheme from another.

Note that (14) approximates (5) in the so-called "conservation form", that is, total differentials are approximated by perfect differences. This allows the use of the approximation in regions where the flow is discontinuous; see Lax and Wendroff (1960).

The methods usually are stable under the Courant-Friedrichs-Lewy condition, which says that the largest radial wave or material speed in a cell must not exceed the numerical signal speed $\Delta\eta/\Delta t$.

In programming methods (a), (b), and (c), the effect of the source term was accounted for in separate steps, that is, Eq. (5) was approximated by starting out with integrating the equation

$$\frac{\partial U}{\partial t}=H, \tag{16}$$

over a half time-step, then continuing by integrating

$$\frac{\partial U}{\partial t}+\frac{\partial F}{\partial\eta}=0, \tag{17}$$

over a full time-step, and finishing off with integrating Eq. (16) over a half time-step. This procedure is convenient in that it allows a very flexible program structure.

Another advantage of this method is that Eq. (16) can often be integrated accurately by itself. For the current problem the analytic solution of (16) is

$$\begin{aligned}
&\varrho(t_n+\tau,\eta_i)=\varrho_i^n,\\
&u(t_n+\tau,\eta_i)-u_0=(u_i^n-u_0)\cos(\kappa\tau)\\
&\qquad+\frac{2\Omega}{\kappa}\left(v_i^n-v_0+\frac{A}{\alpha r\Omega}\sin\hat\eta\right)\sin(\kappa\tau),\\
&v(t_n+\tau,\eta_i)-v_0=\frac{-A}{\alpha r\Omega}\sin\hat\eta\\
&\qquad+\left(v_i^n-v_0+\frac{A}{\alpha r\Omega}\sin\hat\eta\right)\cos(\kappa\tau)-\frac{\kappa}{2\Omega}(u_i^n-u_0)\sin(\kappa\tau).
\end{aligned} \tag{18}$$

Use of such an accurate solution proved essential for maintaining stability in very long runs (over 2000 time-steps). The cause of the instability that arises otherwise is the same that would make a linear first-order algorithm for integrating (16) unstable: instead of choosing $(u(t_n+\tau),\ v(t_n+\tau))$ on the ellipse given by (18), the linearized version will put it on the tangent to that ellipse, thus always leading to an amplification of the disturbance.

The exact solution to the test problem is known reliably only in the steady-state limit. Thus, the numerical methods cannot be evaluated precisely on how well they calculate the time evolution of the flow. We therefore restrict ourselves to testing the methods on the accuracy of the steady state they produce. Note that evaluating the methods on the time required for them to reach the steady-state starting, say, from uniform initial values $(\varrho,u,v)_i^0=(1,u_0,v_0)$, is unfair because the better methods will fare the worst. Because of the periodicity of the flow any transients will persist until they are damped out by the numerical viscosity, which is highest for the least accurate schemes. It seems to us that the fairest test is to use the exact steady state solution itself as the initial-value distribution, and compare how well the various methods preserve it. This constitutes the first test performed. In keeping with the spirit of the methods, the grid values U_i^0 employed are not the point-values of the steady-state but rather the zone-averaged values. Two methods that performed well in this test were applied to the problem with uniform initial values, mainly to determine their "robustness". This constitutes the second test.

The results presented are based on a computational grid of 64 zones; the numerical solutions for all methods are advanced by 1200 time-steps from the exact steady-state with a constant time-step corresponding initially to a global Courant number of 0.5. In this span the fastest moving signals can traverse the computational domain about 10 times. This is a reasonable amount of time to allow the computed solution to adjust towards the steady-state of the difference equations.

Many additional experiments were run with grid sizes of 16–128 zones, a variety of Courant numbers and a maximum number of time steps well over 10,000. These mainly served to check the consistency of our findings or to examine the long-term stability of the numerical solution.

The computed solutions of the first test are displayed in Figs. 1–6, while root-mean-square (rms) errors are listed in Table 1 for three consecutive output times. Since no method is expected to be equally accurate at the shock and away from the shock, 8 points straddling the shock, including 5 points in the decompression region, are excluded from the rms error calculation. The entries in Table 1 therefore mainly indicate the accuracy in the smooth part of the solution; for shock rendition we rely on visual inspection of the figures.

a) The Beam Scheme (B)

The beam scheme, used only in astrophysics, is due to Prendergast (see Sanders and Prendergast, 1974). It exemplifies the "Boltzmann approach," that is, mass and momentum are transported by pseudo-particles with a velocity distribution $f(w)$ designed for numerical convenience. In the beam scheme the velocity distribution is the sum of a number of delta functions (the beams); for the present calculations we used three beams:

$$f(w)=\tfrac{1}{6}\varrho\delta(w-[u-c\sqrt{3}])+\tfrac{2}{3}\varrho\delta(w-u)+\tfrac{1}{6}\varrho\delta(w-[u+c\sqrt{3}]), \tag{19}$$

although the middle beam is not really needed for this isothermal problem. Assuming that the velocity distribution is uniform and constant in each cell during the time-step, we can compute the flux of mass and momentum out of a cell in the positive direction,

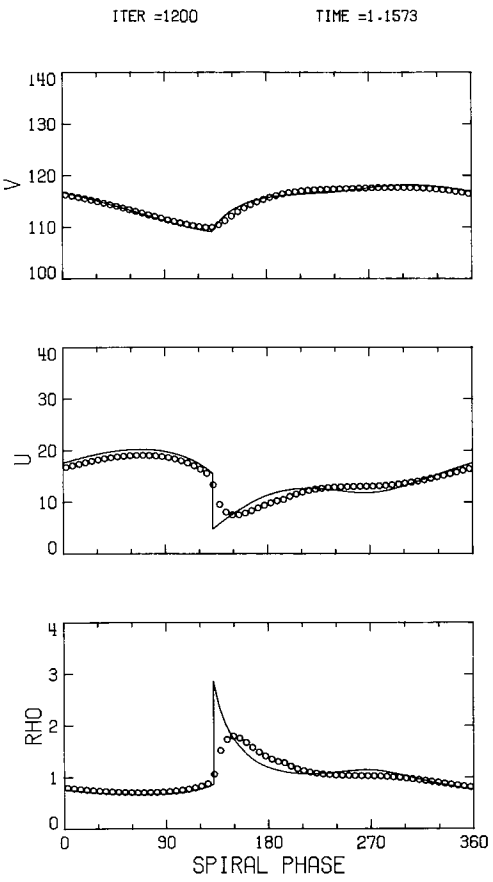

Fig. 1. Numerical results (circles) obtained with the Beam scheme (B) after 1200 time steps, starting from the exact solution (line), zone-averaged

Table 1. Test 1. Rms errors in the smooth region of the numerical solution, obtained with four schemes after 1200 ± 150 time steps with a Courant number of 0.5, using the zone-averaged exact solution as initial-value distribution. Differences in the definition of the Courant number cause differences in the output times t_n. The results at $n=1200$ correspond to those of Figs. 1–6

Scheme	n	t_n	rms error (% of equilibrium value) in ϱ	u	ϱu	v
	1050		7.8	7.6	0.70	0.37
B	1200	1.1573	8.4	7.7	0.77	0.36
	1350		8.0	7.8	0.74	0.37
	1050		3.5	4.0	0.57	0.18
G	1200	1.2266	3.6	4.1	0.40	0.19
	1350		3.6	4.5	0.48	0.19
	1050		0.54	0.42	0.25	0.026
FS2	1200	1.1983	0.62	0.53	0.25	0.020
	1350		0.48	0.46	0.11	0.026
	1050		1.0	0.71	0.58	0.052
MC2	1200	1.1962	1.0	1.0	0.51	0.037
	1350		0.79	0.80	0.32	0.049

$F^+(U)$, and in the negative direction, $F^-(U)$:

$$F^+(U)=\begin{cases} F(U)=\begin{pmatrix}\varrho u\\ \varrho(u^2+c^2)\\ \varrho uv\end{pmatrix} & u\geqq c\sqrt{3}\\ \begin{pmatrix}\frac{1}{6}\varrho(5u+c\sqrt{3})\\ \frac{1}{6}\varrho[4u^2+(u+c\sqrt{3})^2]\\ \frac{1}{6}\varrho v(5u+c\sqrt{3})\end{pmatrix} & 0\leqq u<c\sqrt{3}\\ \begin{pmatrix}\frac{1}{6}\varrho(u+c\sqrt{3})\\ \frac{1}{6}\varrho(u+c\sqrt{3})^2\\ \frac{1}{6}\varrho v(u+c\sqrt{3})\end{pmatrix} & -c\sqrt{3}<u<0\\ 0 & u\leqq -c\sqrt{3}, \end{cases} \tag{20.1}$$

and $F^-(U)$ is obtained from

$$F^-(U)+F^+(U)=F(U). \tag{20.2}$$

The net flux across the cell interface at $\eta_{i+1/2}$, to be used in the scheme (14), is

$$\Phi^n_{i+1/2}=F^+(U^n_i)+F^-(U^n_{i+1}). \tag{21}$$

That Eq. (21) leads to upwind differencing becomes clear when we write down the central difference of Φ needed in scheme (14):

$$\Phi^n_{i+1/2}-\Phi^n_{i-1/2}=(F^+)^n_i-(F^+)^n_{i-1}+(F^-)^n_{i+1}-(F^-)^n_i. \tag{22}$$

Thus, the flux difference of the forward/backward moving material is biased in the backward/forward direction.

The local stability condition of a beam scheme with velocity dispersion $\pm\mu c$, $\mu\geqq1$, is

$$\frac{\Delta t}{\Delta\eta}(|u|+\mu c)_i\leqq1. \tag{23}$$

The beam scheme turns out to coincide, for $\mu=1$, with the "flux-vector splitting" method of Steger and Warming (1978), derived without reference to a velocity distribution. The term "splitting" refers to the splitting of F into F^+ and F^-. Figure 1 displays the results obtained with the beam scheme.

b) Godunov's Method (G)

In this method, designed by Godunov (1959), the interaction of a pair of cells at their interface is assumed to take place through discrete finite-amplitude waves rather than material beams. This yields a more accurate scheme. It has less diffusion than the beam scheme, yet still avoids non-physical oscillations in the vicinity of shock waves.

Assuming that at $t=t_n$ the distributions in each cell are uniform, we find discontinuities at the cell interfaces; these will be resolved instantaneously through shock waves and/or rarefaction waves, and (not for an isothermal gas) a contact discontinuity. Let $\hat{U}^n_{i+1/2}$ be the state remaining at $\eta_{i+1/2}$ after the resolution; Godunov's method incorporates

$$\Phi^n_{i+1/2}=F(\hat{U}^n_{i+1/2}). \tag{24}$$

It is easily seen that

$$(\delta F^-)^n_{i+1/2}\equiv\Phi^n_{i+1/2}-F^n_i, \tag{25}$$

and

$$(\delta F^+)^n_{i+1/2}\equiv F^n_{i+1}-\Phi^n_{i+1/2} \tag{26}$$

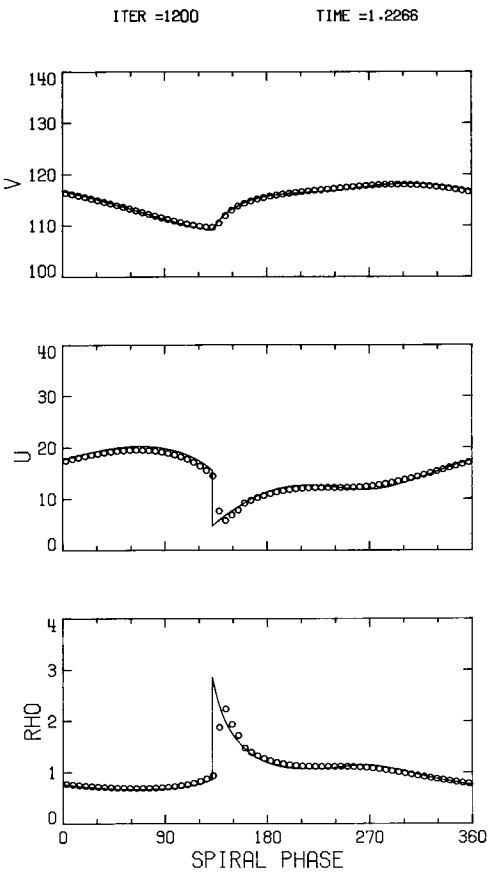

Fig. 2. Results of Godunov's method (G) after 1200 time steps

are the flux differences across the waves moving backward and forward from $\eta_{i+1/2}$, respectively. We have

$$\Phi^n_{i+1/2}-\Phi^n_{i-1/2}=\Phi^n_{i+1/2}-F^n_i+F^n_i-\Phi^n_{i-1/2}$$
$$=(\delta F^-)^n_{i+1/2}+(\delta F^+)^n_{i-1/2}, \tag{27}$$

showing the upwind character of this method. The splitting evident in (25), (26), that is

$$\Delta_{i+1/2}F^n\equiv F^n_{i+1}-F^n_i=(\delta F^-)^n_{i+1/2}+(\delta F^+)^n_{i+1/2}, \tag{28}$$

is called flux-difference splitting, as opposed to the flux-vector splitting of Eq. (20.2).

The computation of $\hat{U}^n_{i+1/2}$ is the bottle-neck of Godunov's method, and a multitude of variations have been proposed based on employing a suitable approximation of $\hat{U}^n_{i+1/2}$; see Harten et al. (1981). In our computations we used an iterative procedure to find $\hat{U}^n_{i+1/2}$, which we stopped just short of convergence. For a compilation of the formulas and iteration procedures to resolve an arbitrary discontinuity (Riemann's initial-value problem) the reader may consult, e.g., Chorin (1976) or Van Leer (1979).

Note that in the supersonic region of the flow $U^n_{i+1/2}$ equals either U^n_i or U^u_{i+1} (the scheme becomes one-sided), so no full Riemann solution is needed. It therefore pays to test whether the flow is sub- or supersonic. The local stability condition for Godunov's method is

$$\frac{\Delta t}{\Delta\eta}W_{i+1/2}\leqq 1, \tag{29}$$

when $W_{i+1/2}$ is the largest wave speed occurring in the solution of the Riemann problem at $\eta_{i+1/2}$. The results obtained with the Godunov method are displayed in Fig. 2.

c) Second-order Flux-splitting Method (FS2)

Any first-order upwind-differencing method can be changed into a second-order method by first advancing the cell-boundary values, to be used in the numerical flux function, and the source term, to the intermediate time level $t_{n+1/2}$. In obtaining these values, *the interaction between cells can be fully ignored.* This observation, due to Hancock (1980), has led to a drastic simplification of second-order upwind schemes since these first were formulated by Van Leer (1979).

We choose Q to be a vector of (not necessarily conserved) quantities describing the state of the gas, in particular:

$$Q=\begin{pmatrix}\varrho\\ u\\ v\end{pmatrix}=\begin{pmatrix}\varrho\\ \boldsymbol{q}\end{pmatrix}. \tag{30}$$

We then assume that the initial values for Q form a piecewise linear distribution:

$$Q^n(\eta)=Q^n_i+(\eta-\eta_i)\frac{(\delta Q)^n_i}{\Delta\eta}\qquad \eta_{i-1/2}<\eta<\eta_{i+1/2} \tag{31.1}$$

with

$$(\delta q)^n_i=c\cdot\mathrm{ave}\left(\frac{q^n_{i+1}-q^n_i}{c},\frac{q^n_i-q^n_{i-1}}{c}\right) \tag{31.2}$$

and

$$(\delta\varrho)^n_i=\varrho^n_i\cdot\mathrm{ave}\left(2\frac{(\varrho^n_{i+1}-\varrho^n_i)}{(\varrho^n_{i+1}+\varrho^n_i)},2\frac{(\varrho^n_i-\varrho^n_{i-1})}{(\varrho^n_i+\varrho^n_{i+1})}\right), \tag{31.3}$$

where ave (a, b) is an averaging procedure to be specified later. The formulation in Eq. (31.3) guarantees positivity for ϱ when substituted in Eq. (31.1). Thus we have

$$\left(\frac{\partial Q}{\partial\eta}\right)^n_i=\frac{(\delta\Omega)^n_i}{\Delta\eta}, \tag{32}$$

allowing us to calculate $(\partial Q/\partial t)^n_i$ from the appropriate modification of Eq. (5). The cell averages are now advanced to $t_{n+1/2}$, and boundary values are calculated [the source terms have already been advanced, by Eq. (16)]:

$$Q^{n+1/2}_i=Q^n_i+\frac{\Delta t}{2}\left(\frac{\partial Q}{\partial t}\right)^n_i, \tag{33.1}$$

$$Q^{n+1/2}_{(i\pm1/2)_\mp}=Q^{n+1/2}_i\pm\frac{\Delta\eta}{2}(\delta Q)^n_i, \tag{33.2}$$

$$U^{n+1/2}_{(i\pm1/2)_\mp}=U(Q^{n+1/2}_{(i\pm1/2)_\mp}). \tag{33.3}$$

The time-centered fluxes at cell boundary $i\pm\frac{1}{2}$ can now be computed from $U^{n+1/2}_{(i\pm1/2)_-}$ and $U^{n+1/2}_{(i\pm1/2)_+}$ by any upwind-biased numerical flux formula, such as used in B or G. Here we use a formula, due to Van Leer (1981b), based on flux-vector splitting and therefore related to the flux in B; however, no particular

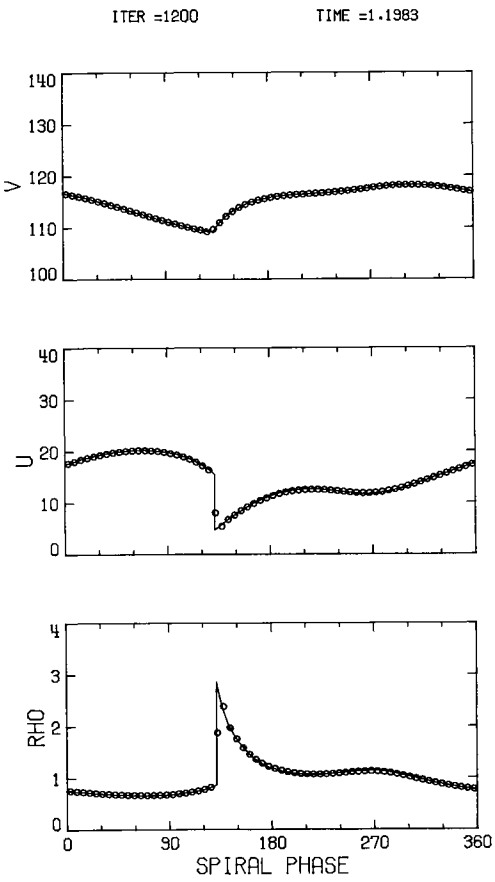

Fig. 3. Results of the second-order flux-splitting method (FS2) after 1200 time steps

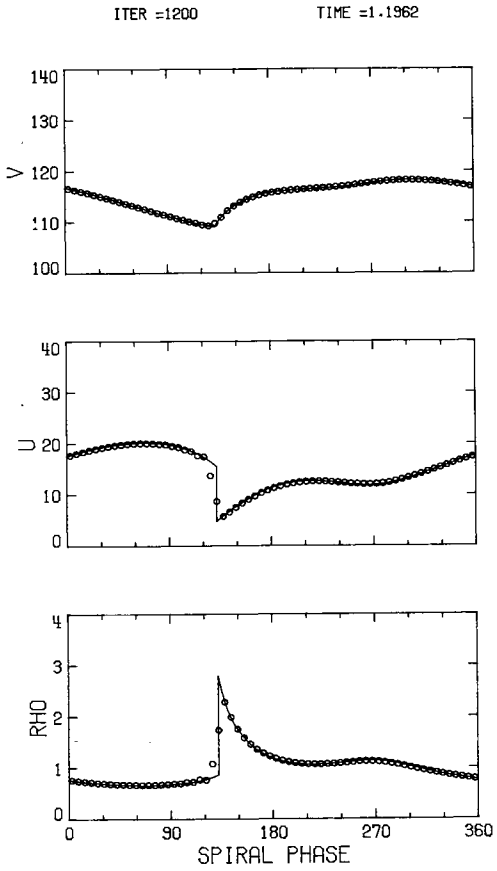

Fig. 4. Results of the second-order MacCormack method (MC2) after 1200 time steps

velocity distribution is used in its derivation. The forward and backward fluxes of mass and momentum are defined according to

$$F^+(U)=\begin{cases} F(U) & u\leqq c \\ \begin{pmatrix} \frac{\varrho}{4c}(u+c)^2 \\ \frac{\varrho}{2}(u+c)^2 \\ \frac{\varrho}{4c}(u+c)^2 v \end{pmatrix} & |u|<c, \\ 0 & u\leqq -c \end{cases} \tag{34}$$

and (20.2), while Φ again is given by (21). The split flux (34) is smoother than (20), having continuous first derivatives. Furthermore, the reduced mass flux in (34) relative to (20.1) results in a reduced numerical diffusion, making the numerical results of (34) alone come very close to the results of Godunov's flux (24) for the present test problem. The local stability condition is a combination of (23) with $\mu=1$, and

$$\frac{\Delta t}{\Delta \eta}\, c_i \leqq \frac{1}{2}. \tag{35}$$

In FS2 we then use

$$\Phi^{n+1/2}_{i+1/2}=F^+(U^{n+1/2}_{(i+1/2)-})+F^-(U^{n+1/2}_{(i+1/2)+}), \tag{36}$$

with $F^+(U)$ and $F^-(U)$ given by (34) and (20.2).

The function ave(a,b) is chosen such that it tends to $\frac{1}{2}(a+b)$ if a and b are subsequent finite differences of a smooth solution, but tends to the smallest value where the solution is not smooth (see Van Leer, 1977). We specifically chose

$$\text{ave}(a,b)=\frac{(b^2+\varepsilon^2)a+(a^2+\varepsilon^2)b}{a^2+b^2+2\varepsilon^2}, \tag{37}$$

where ε^2 is a small non-vanishing bias of the order $O((\Delta\eta)^3)$. This type of averaging prevents central differencing across a discontinuity in the solution or in its first derivative, which would lead to numerical oscillations. The bias prevents the undesirable clipping of a smooth extremum (see Sect. e) but otherwise has negligible influence. In the actual computations we used $\varepsilon^2=0.008$, but the results are not very sensitive to the precise value of ε^2. The results obtained with the second-order flux-splitting method are displayed in Fig. 3.

d) MacCormack Method (MC2)

This finite-difference method developed by MacCormack (1969) has been used widely in aerodynamics. On even time-steps it uses a forward predictor step which determines provisional values at t_{n+1}

$$\bar{U}^{n+1}_i=U^n_i-\frac{\Delta t}{\Delta y}(F^n_{i+1}-F^n_i)+\Delta t H^n_i, \tag{38}$$

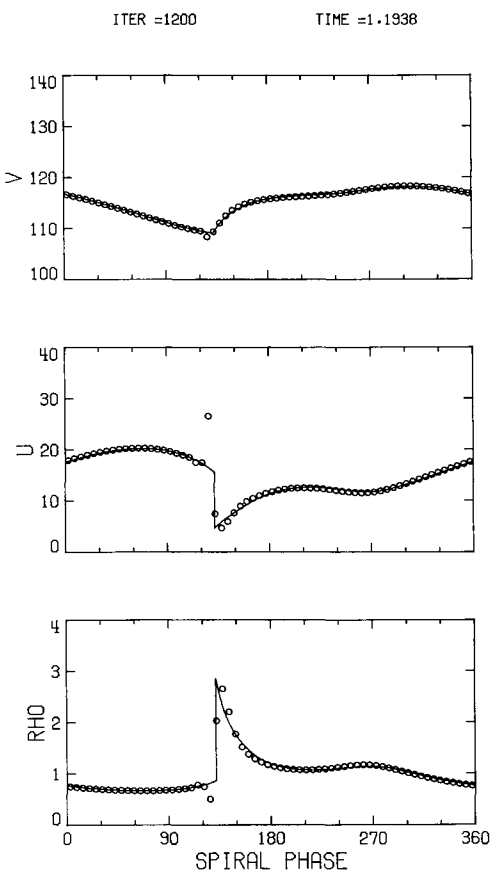

Fig. 5. Results of SHASTA after 1200 time steps as given by Eqs. (43)–(45), with the limiter *off*

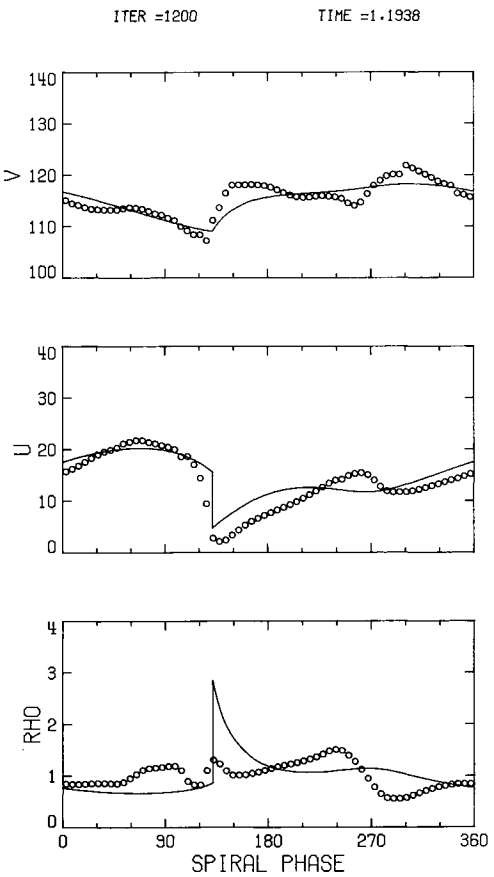

Fig. 6. Results of SHASTA after 1200 time steps, with the limiter *on*

followed by a backward corrector step which determines the final values at t_{n+1}

$$U_i^{n+1}=\tfrac{1}{2}\left[U_i^n+\bar{U}_i^{n+1}-\frac{\Delta t}{\Delta\eta}(\bar{F}_i^{n+1}-\bar{F}_{i+1}^{n+1})+\Delta t\,\bar{H}_i^{n+1}\right]. \tag{39}$$

The corrector step corresponds to inserting

$$\Phi_{i+1/2}^{n+1/2}\equiv\tfrac{1}{2}(F_{i+1}^n+\bar{F}_i^{n+1}), \tag{40.1}$$

$$H_i^{n+1/2}\equiv\tfrac{1}{2}(H_i^n+\bar{H}_i^{n+1}), \tag{40.2}$$

into (14). On odd time-steps a backward predictor is employed with a forward corrector. This method is formally second-order accurate in both space and time.

Although MacCormack's method is slightly dissipative, an explicit smoothing term had to be added in order to control nonlinear instabilities in the test problem. This was implemented by adding the term

$$D_i^n=b\frac{\Delta t}{\Delta\eta}[v_{i+1/2}^n(U_{i+1}^n-U_i^n)-v_{i-1/2}^n(U_i^n-U_{i-1}^n)] \tag{41}$$

to the right-hand side of Eq. (39). The coefficient b is an adjustable constant of order unity; the choice

$$v_{i+1/2}^n=|u_{i+1}^n-u_i^n|, \tag{42}$$

for the artificial-viscosity coefficient was made on the basis of Liebovitch's (1978) experience with MacCormack's method in two-dimensional calculations. Figure 4 displays the results obtained with the second order MacCormack method [provided by T. A. Zang following the calculations of Zang and Hussaini (1980)].

e) Flux-corrected Transport Methods

We tested the following Flux-Corrected Transport (FCT) methods: SHASTA, Boris and Book (1973), and Phoenical SHASTA, Book et al. (1975) in the formulation given below, and also the low-phase-error routine of Boris (1976). These methods are known to perform well for flows with strong transient waves; we verified this with our programs.

The simplest FCT methods can be implemented by adding to the scheme of Lax and Wendroff (1960) or MacCormack (38, 39) the strong artificial diffusion

$$D_i^n=\tfrac{1}{8}(U_{i+1}^n-2U_i^n+U_{i-1}^n), \tag{43}$$

thus suppressing numerical oscillations for a local Courant number $<\frac{1}{2}\sqrt{3}$, and then taking the diffusion away in the so-called anti-diffusion step, without creating new oscillations:

$$\tilde{U}_i^{n+1}=\tilde{U}_i^{n+1}-\tfrac{1}{8}\{(\tilde{U}_{i+1}^{n+1}-\tilde{U}_i^{n+1})_{\lim} -(\tilde{U}_i^{n+1}-\tilde{U}_{i-1}^{n+1})_{\lim}\} \quad \text{(SHASTA)}, \tag{44}$$

$$\tilde{U}_i^{n+1}=\tilde{U}_i^{n+1}-\tfrac{1}{8}(U_{i+1}^{n+1}-U_i^{n+1})_{\lim} -(U_i^{n+1}-U_{i-1}^{n+1})_{\lim} \quad \text{(Phoenical SHASTA)}. \tag{45}$$

Table 2. Test 2. Rms errors in the smooth region of the numerical solution, obtained with two schemes (FS2 and MC2) after 2400 time steps with a Courant number of 0.5, using uniform initial values

			rms errors (% of equilibrium value) in			
Scheme	n	t_n	ϱ	u	ϱu	v
FS2	2400	2.3946	2.4	4.8	7.0	0.20
MC2	2400	2.3934	3.1	3.9	7.0	0.19

Here U^{n+1} is the updated solution without the extra diffusion, $\tilde{U}^{n+1}$ is the diffused updated solution and $\tilde{\tilde{U}}^{n+1}$ is the final result. The limiter works as follows:

$$(U_{i+1}-U_i)_{\lim}=$$
$$\begin{cases}\min\{8|U_i-U_{i-1}|,|U_{i+1}-U_i|,8|U_{i+2}-U_{i+1}|\}\operatorname{sgn}(U_{i+1}-U_i)\\ \quad \text{if} \quad \operatorname{sgn}(U_i-U_{i-1})=\operatorname{sgn}(U_{i+1}-U_i)=\operatorname{sgn}(U_{i+2}-U_{i+1}),\\ 0 \text{ otherwise.}\end{cases} \tag{46}$$

In integrating our periodic test-problem starting from the stationary solution, we find that the limiter tries to give each extremum in ϱ, ϱu, and ϱv a flat top. This generates waves that keep running around, pile up and ultimately destroy any resemblance of the numerical results to the stationary solution.

Figures 5 and 6 show the results obtained with the Lax-Wendroff-based SHASTA. In order to illustrate the effect of the limiter, we first display the results generated with the limiter switched off in Fig. 5; these have reached a steady-state to the same degree as methods a)–d). Note the pre-shock spatial oscillations; without the limiter the scheme does not preserve monotonicity. Figure 6 shows the results generated with the limiter switched on. These are far from steady, and there is no numerical steady state to which they can converge. The other FCT methods tested behaved similarly.

A recent investigation of FCT methods by Zalesak (1981) suggests that it is not the form of the limiter that causes the problems, but the fact that it is placed in the final step of these methods, and that it acts independently on each of the conserved quantities, without synchronization.

It must be mentioned that the method FS2 also contains a limiter, albeit much more gently than (46), in the form of the averaging function (37). To our knowledge this does not hamper the convergence of the numerical solution in any way. However, divergence was observed with an earlier version of the formula, used by Van Leer (1977, 1979):

$$\text{ave}\,(a,b)=\frac{|b|a+|a|b}{|a|+|b|}. \tag{47}$$

This causes clipping of an extremum, since

$$\text{ave}\,(a,b)=0 \quad \text{if} \quad \operatorname{sgn} a=-\operatorname{sgn} b. \tag{A6}$$

In the present test problem, use of (47) in an FCT-like formulation of scheme FS2 caused a slowly growing travelling disturbance, invisible after 1200 time steps but a clear lump after 3600 time steps.

IV. Results

Method B a) is clearly singled out for its strong numerical diffusion. The computed shock appears as a shadow of the real one. The density maximum is severely underestimated and is displaced downstream by several zones. The corresponding values of u are barely subsonic. In the smooth region of the flow only the most general features are represented.

The displacement of the shock in the downstream direction is typical of upstream-differencing methods; the supersonic zone immediately before the shock is not influenced by the downstream subsonic region.

Method G b), although of the same order of accuracy, produces a substantial improvement. The density peak is better represented, the shock is much narrower, although still displaced downstream, and the smooth part of the solution is also closer to the truth.

Note the change in the numerical derivative of u and ϱ across the sonic point near a spiral phase of 160°. The downstream subsonic region cannot numerically influence the upstream supersonic region by sound waves moving upstream, while the numerical diffusion across those waves vanishes in the sonic point. The solutions on either side therefore are not strongly coupled. We may also say that the rapid variation of the scheme's truncation error shows up on this relatively coarse grid.

The very best results are produced by FS2 c). Its sharp, oscillation-free rendition of the shock shows the efficiency of applying dissipation through the averaging function (37). The rms entries in Table 1 are uniformly the lowest in all columns.

The virtues of method MC2 d) are evident in the smooth region of the flow. The shock is reasonably narrow and in the right place, with minor oscillations in front. The free parameter b in the numerical dissipation (41) was given the value 1.0; values in the range 0.2–2.0 produce acceptable results. No "fine-tuning" was attempted. Nonetheless, the fact that MC2 needs an adjustable parameter must be considered as a disadvantage of this method.

For methods a)–d) the rms error in the product ϱu, a quantity that becomes uniform in the steady state, is considerably smaller than the error in each of its factors. The errors in v, which by itself varies only a few percent in the solution, are again much smaller. The numbers in Table 1 are representative in that they do not change significantly when the iterations in time are continued far beyond 1200. Because of their poor quality, the FCT results have not been included in Table 1.

The methods, FS2 and MC2 were selected for the second test, to approch the steady state from uniform initial values. Neither of the schemes developed stability problems. The rms errors after 2400 time steps are listed in Table 2. They do not differ much from scheme to scheme and are substantially larger than the steady-state errors.

V. Conclusions and Recommendations

From the results of the first test a number of conclusions can be drawn about the spatial accuracy of the methods considered. The following three are quite firm.

1. The best second-order method (FS2) outperforms the best first-order method (G) by a huge margin, on a grid with resolution comparable to what is feasible in two-dimensional calculations. The second-order method, which includes a predictor step and a monotonicity algorithm, requires less than twice the computation time of the first-order method. A comparison of the rms errors suggests that the first-order method would require a six times finer

grid to match the accuracy of the second-order method. Thus, the extra programming effort reduces the total computational cost by an order of magnitude for the present mesh size. This holds even more strongly for computations on finer grids or in more dimensions.

2. Monotonicity algorithms for second-order methods have really come of age. The averaging procedure in FS2 that replaces central differencing is a simple and effective implementation of the ideas of artificial dissipation, filtering and limiting that have been explored during the last decade. The performance of FS2 derives to a great extent from the availability of this algorithm.

3. The second-order central-differencing scheme (MC2), while sufficiently accurate in rendering the smooth part of a solution, cannot compete with the second-order upwind-differencing method (FS2) in shock representation. The culprit seems to be the kind of artificial-dissipation term commonly used with this scheme. Better results can be expected from the use of terms derived from the scheme's truncation error (Klopfer and McRae, 1981); the derivation, however, is cumbersome.

The results of the second test show that the schemes selected could handle the more severe transients without heavily damping them. Firm conclusions about the accuracy of these schemes applied to transient phenomena should not be given, since the solutions obtained cannot be calibrated with a time-dependent exact solution.

One transient problem with known exact solution is the Riemann problem, describing the break-up of a discontinuity between two arbitrary uniform states. Comparative tests of twelve schemes on the basis of such a problem were carried out by Sod (1978). Among the correctly programmed schemes are central-difference schemes, including MC2. The results obtained for the same Riemann problem by Van Leer (1979) with an earlier version of FS2 are more accurate than any of the results of Sod, indicating the same hierarchy of accuracy as established in the present paper.

Our test results do not indicate how to proceed in order to reach a steady state that is not known a priori, with the minimum number of time steps. This was not the goal of the present paper. One possible strategy might be to start out with a first-order scheme while slowly turning on the source terms, and then finish off with a higher-order scheme.

In summary, on the basis of the present test results we may recommend the second-order upwind-differencing method FS2 for solving flow problems involving a shock of importance. For smoother problems, MC2 may do as good a job. One advantage of MC2 is that it has been extensively experimented within multi-dimensional computations, while the multi-dimensional experience with FS2 is still limited.

Acknowledgements. We are much indebted to T. A. Zang and M. Y. Hussaini for proposing the specific test problem, providing the MacCormack method, and for contributing to the manuscript.

References

Book, D.L., Boris, J.P., Hain, K.H.: 1975, *J. Computational Phys.* **18**, 248

Boris, J.P.: 1976, U. S. Naval Research Laboratory Memorandum Report 3237

Boris, J.P., Book, D.L.: 1973, *J. Computational Phys.* **11**, 38

Chorin, A.J.: 1976, *J. Computational Phys.* **22**, 517

Godunov, S.K.: 1959, *Mat. Sb.* **47**, 271 (also see Cornell Aeronautical Lab. Transl.)

Hancock, S.L.: 1980 (private communication)

Harten, A., Lax, R.D., Van Leer, B.: 1981, ICASE Report (in preparation)

Klopfer, G.H., McRae, D.S.: 1981, Proc. of AIAA Computational Fluid Dynamics Conference, Palo Alto, CA, 317

Lax, P.D., Wendroff, B.: 1960, *Comm. Pure Appl. Math.* **13**, 217

Liebovitch, L.S.: 1978, Ph. D. Thesis, Harvard University

Lin, C.C., Lau, Y.Y.: 1979, *Studies in Appl. Math.* **60**, 97

Lin, C.C., Shu, F.H.: 1964, *Astrophys. J.* **140**, 646

Lin, C.C., Shu, F.H.: 1966, *Proc. Natl. Acad. Sci. USA* **55**, 229

MacCormack, R.W.: 1969, AIAA Paper No. 69–354

Roberts, W.W.: 1969, *Astrophys. J.* **158**, 123

Sanders, R.H., Prendergast, K.H.: 1974, *Astrophys. J.* **188**, 489

Shu, F.H., Milione, V., Roberts, W.W.: 1973, *Astrophys. J.* **183**, 819

Sod, G.A.: 1978, *J. Computational Phys.* **27**, 1

Steger, J., Warming, R.: 1978, *J. Computational Phys.* **40**, 263

Toomre, A.: 1977, *Ann. Rev. Astron. Astrophys.* **15**, 437

Van Leer, B.: 1977, *J. Computational Phys.* **23**, 276

Van Leer, B.: 1979, *J. Computational Phys.* **32**, 101

Van Leer, B.: 1981a, ICASE Report No. 81–11

Van Leer, B.: 1981b, ICASE Report (in preparation)

Woodward, P.R.: 1975, *Astrophys. J.* **195**, 61

Zalesak, S.T.: 1981 (private communication)

Zang, T.A., Hussaini, M.Y.: 1980, Proc. 7th Intl. Conf. on Numerical Methods in Fluid Dynamics: *Lecture Notes in Physics*, Springer, Berlin, Heidelberg, New York