INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 18, 273–294 (1994)

A CONSERVATIVE AND SHAPE-PRESERVING SEMI-LAGRANGIAN METHOD FOR THE SOLUTION OF THE SHALLOW WATER EQUATIONS

P. GARCIA-NAVARRO* AND A. PRIESTLEY

Department of Mathematics, PO Box 220, University of Reading, Whiteknights, Reading, U.K.

SUMMARY

Semi-Lagrangian methods are now perhaps the most widely researched algorithms in connection with atmospheric flow simulation codes. In order to investigate their applicability to hydraulic problems, cubic Hermite polynomials are used as the interpolant technique. The main advantage of such an approach is the use of information from only two points. The derivatives are calculated and limited so as to produce a shape-preserving solution. The lack of conservation of semi-Lagrangian methods, however, is widely regarded as a serious disadvantage for hydraulic studies, where non-linear problems in which shocks may develop are often encountered. In this work we describe how to make the scheme conservative using an FCT approach. The method proposed does not guarantee an unconditional shock-capturing ability but is able to correctly reproduce the discontinuous flows common in open channel simulation without any shock-fitting algorithm. It is a cheap way to improve existing 1D semi-Lagrangian codes and allows stable calculations beyond the usual CFL limits. A basic semi-Lagrangian method is presented that provides excellent results for a linear problem; the new techniques allow us to tackle non-linear cases without unduly degrading the accuracy for the simpler problems. Two one-dimensional hydraulic problems are used as test cases, water hammer and dam break. In the latter case, because of the non-linearity, special care is needed with the low-order solution and we show the advantages of using Leveque's large-time step version of Roe's scheme for this purpose.

KEY WORDS Method of characteristics Polynomial interpolation Monotonicity Recovery of conservation

1. INTRODUCTION

Semi-Lagrangian-based methods, or interpolation methods as they are sometimes called in the hydraulics literature, have proved very successful in computational fluid dynamics (CFD), mainly in connection with atmospheric flow prediction where they have gained widespread acceptance.[1] They can be described as a technique using a fixed grid that essentially combines the method of characteristics with a suitable interpolating procedure. The basic idea of following characteristics backwards in time in order to pick the correct information from the past comes from very early works in CFD.[2] Semi-Lagrangian methods can be distinguished from one another by the interpolant used at the foot of the characteristic.

Traditionally, the finite difference version of the method of characteristics has been applied with success to hydraulic transients in pipes[3] and rivers.[4,5] It is known, nevertheless, that the use of a linear interpolation between two computational points sometimes produces an excessive

* On leave from the University of Zaragoza, Zaragoza, Spain.

0271–2091/94/030273–22$16.00

Received May 1993
Revised September 1993

amount of attenuation. The numerical damping can be reduced by using a more refined interpolation algorithm which will determine the spatial accuracy of the scheme. In recent years various higher-order interpolants have been proposed. Some of them use four or more computational points to construct cubic, which can actually be shown to be equivalent to the finite element Lagrange Galerkin method[6] with linear elements,[7] or quintic polynomials. As an alternative approach, the use of no more than two points is possible if the spatial derivatives of the polynomial at these points are supplied in order to provide enough information to determine the polynomial.

Higher-order interpolations may lead, however, to spurious numerical oscillations in regions of steep gradients of the interpolated variables. Special limiting or shape-preserving techniques are then required.[8] Moreover, semi-Lagrangian methods are not conservative, except in trivial cases, and hence are inefficient and inaccurate when discontinuities occur in the solution. The usual way to cope with this disadvantage has been the addition of a shock-fitting algorithm connecting the regions of smooth flow.[9]

In this paper we shall concern ourselves with the performance of shape-preserving Hermite cubic polynomials as the interpolant technique when implementing a semi-Lagrangian scheme to solve the 1D shallow water equations. We will also consider the applicability and limitations of new ways of coping with the lack of the important property of conservation.

Below we shall briefly describe the semi-Lagrangian algorithm and the implementation of shape-preserving solutions using Hermite polynomials. In Section 2 a technique to recover conservation will be introduced for the scalar case as well as for systems of equations. It will be pointed out that some difficulties can be met when applying it to a system of equations. In Section 3 an extension will be proposed to overcome this difficulty.

To demonstrate its effectiveness and to facilitate this study, two tests problems from the hydraulics literature have been selected and several numerical results are shown.

1.1 Test Problems

Case 1: dam break flow. Even though the strategy to recover conservation in semi-Lagrangian schemes is not intended to produce a method able to cope with strong discontinuities, as in other shock-capturing methods, and our primary interests are in river and pipe flows, where discontinuities are generally weaker, the idealized dam break problem was chosen because it is a classical example of non-linear flow with shocks to test conservation in numerical schemes and at the same time has an analytical solution.

This problem is generated by the one-dimensional shallow water equations given by

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0, \qquad \frac{\partial Q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{Q^2}{A} + gI_1\right) = 0, \tag{1}$$

where A is the wetted cross-sectional area, Q is the discharge and I_1 represents a hydrostatic pressure force term. For the ideal case of a flat, frictionless channel of unit width and rectangular cross-section we have $A = h$, h being the water depth, and $I_1 = \frac{1}{2}gh^2$, g being the acceleration due to gravity.

The initial conditions are

$$h(x, 0) = \begin{cases} h_L & \text{if } x \leqslant L/2, \\ h_R & \text{if } x > L/2, \end{cases} \qquad Q(x, 0) = 0.$$

Calculation times were used so as to avoid interaction with the extremities of the channel. The boundary conditions are then trivial.

Case 2: the water hammer problem. The linearized water hammer problem, owing to its popularity as an example in the related literature, was selected as a means to compare the performance of the monotone Hermite cubic interpolation against other proposed semi-Lagrangian schemes. It was also used to determine the extent of the effect caused by the recovery of conservation. Being a linear and simplified problem, it was suitable in order to focus attention on the adaptation of the algorithm to systems of equations. Following the dimensionless formulation of Sibetheros and Holley,[10] for instance, this second test problem deals with the solution of the linear system of equations

$$\frac{\partial H}{\partial t} + a\frac{\partial V}{\partial x} = 0, \qquad \frac{\partial V}{\partial t} + a\frac{\partial H}{\partial x} = 0,$$

with the initial conditions

$$H(x, 0) = 0, \qquad V(x, 0) = 1$$

and the boundary conditions

$$H(0, t) = 0, \qquad V(L, t) = 0,$$

where

$$a = \sqrt{(gh_0)}.$$

The dimensionless variables are

$$V = v/v_0, \qquad H = \frac{g(h - h_0)}{av_0},$$

where v is the velocity, h is the specific head and subscripts zero refer to the undisturbed values.

2. BASIC SEMI-LAGRANGIAN SCHEMES

In this section we review the semi-Lagrangian solution to the scalar problem

$$u_t + \mathbf{a}(\mathbf{x}, t) \cdot \nabla u = 0,$$

which describes the advection of $u(\mathbf{x}, t)$. The invariance of a scalar quantity

$$u(\mathbf{x}, t) = u(\mathbf{x}_0, t_0)$$

along a trajectory

$$\mathbf{x}(t) = \mathbf{x}_0(t_0) + \int_{t_0}^{t} \mathbf{a}(\mathbf{x}, \tau)\, d\tau \tag{2}$$

is common to a wide variety of fluid dynamics topics. The aim is to obtain a good approximation of the function $u(\mathbf{x}, t)$ at all the $\mathbf{x}_i$-points of a fixed discrete grid, assuming that u and $\mathbf{a}$ are known everywhere in the grid at an earlier time t_0.

In general, two distinct steps are involved. The first step determines the departure points $\mathbf{x}_0$ of the trajectories arriving at $\mathbf{x}_i$ from the past time through approximate solutions of (2). The

second step is concerned with the way of calculating the value of u at $\mathbf{x}_0$, which in general will not coincide with a grid point, i.e. the way to interpolate u at $\mathbf{x}_0$.

At this stage it is worth stressing a first and important advantage offered by the semi-Lagrangian approach. The usual Courant–Friedrichs–Lewy (CFL) restriction for explicit schemes[2] is no longer a limitation for the stability of the semi-Lagrangian method. Instead, it is replaced by a weaker condition which relates the stability of the resulting scheme to its variability along trajectories and which, for sufficiently smooth flows, permits CFL numbers greatly exceeding unity, thus saving computational effort and improving accuracy away from shocks or regions of strong gradients in the viscous case.[8] Another important advantage of the semi-Lagrangian technique, which we do not make use of in this paper, is the fact that it is genuinely multidimensional.

The first of the two steps referred to above can be achieved in principle by means of any ODE solver.[11] In the particular case of linear advection the exact trajectories are known and used, rendering this step trivial. In the case of a non-linear problem, e.g. the complete shallow water equations, the characteristics will not be straight lines, but Euler's forward difference method can be used giving $O(\Delta t)$ time accuracy. This can be increased to $O(\Delta t^2)$ if an iterative solution of the implicit midpoint rule, common in the meterology literature, is used instead.[1] Unfortunately, that will still not be accurate enough in the neighbourhood of strong shocks or discontinuities. For these problems we cannot achieve any improvement by using increasingly higher-order-accurate methods. A simple way to see how the more classical techniques such as Runge–Kutta fail is to consider a dam break problem where the downstream side is a dry bed and to calculate what happens to the trajectories in this extreme case. Clearly, some sort of implicitness would be beneficial to solve what is now a stiff ODE. A definite improvement is obtained in that case if some forward-in-time information on the slopes of the trajectories is introduced in the Euler procedure. It introduces a certain implicitness and is still efficient in smooth regions. Backward difference formulae are often used to solve stiff ODE systems[11] and may have something to offer here as well. This question is something we hope to return to in a later paper.

Having found a departure point for the trajectory, an adequate interpolation algorithm is necessary and the question of which one is the best suited remains unanswered.

Cubic interpolation seems to have become the most popular in the context of semi-Lagrangian schemes, being more accurate than linear interpolation (which in fact can be recast as a first-order upwind difference method) and less dispersive than quadratic interpolation (which is the equivalent of the second-order Lax–Wendroff explicit scheme).[12] Quintic interpolations have also been proposed but have not gained the same widespread use as cubic ones and will not be considered in this work.

Holly and Preissmann[13] showed that using more than two points to construct a cubic polynomial introduced excessive numerical error due to the physical displacement of the points considered. They constructed cubic and quintic Hermite interpolation polynomials using only two computational points, with the extra information needed coming from the derivatives at those points. However, an auxiliary problem for the first derivative had to be solved. To avoid that difficulty, Schohl and Holly[12] proposed the use of cubic spline interpolation and concluded that the two schemes are of similar accuracy for a contaminant advection problem.

Sibetheros and Holley[10] compared the performances of various types of cubic spline interpolations for a linearized water hammer problem and achieved monotonicity by using a taut cubic spline polynomial which proved adequate for that test case.

Rasch and Williamson[14] were concerned with shape-preserving high-order interpolants as a correct way to improve the already advantageous semi-Lagrangian methods. Monotonicity was

introduced via constraints or restrictions on the derivative estimates at the endpoints of an interval. This approach is used in the present work. Hermite cubic polynomials have been chosen for their simplicity, accuracy and the important advantage of allowing the calculation of the derivatives from the solution itself.

A Hermite cubic polynomial used to provide the interpolated value of a function $f(x)$ defined in a discrete mesh $\{x_i, i = 1, N\}$, $\Delta x_i = x_{i+1} - x_i$, at a point x_p, $x_i \leq x_p \leq x_{i+1}$, can be expressed as

$$p(x_p) = c_1(x_p - x_i)^3 + c_2(x_p - x_i)^2 + c_3(x_p - x_i) + c_4,$$

where the coefficients are

$$c_1 = \frac{d_{i+1} + d_i - 2\Delta_i}{\Delta x_i^2}, \qquad c_2 = \frac{-d_{i+1} - 2d_i + 3\Delta_i}{\Delta x_i}, \qquad c_3 = d_i, \qquad c_4 = f(x_i) = f_i.$$

These coefficients are functions of the discrete slopes Δ_i defined as

$$\Delta_i = \frac{f_{i+1} - f_i}{x_{i+1} - x_i}$$

and of the space derivatives of f at the nodes, d_i, which can be estimated[15] in the case of a uniform mesh spacing Δx by

$$d_i = \frac{-\Delta_{i-2} + 7\Delta_{i-1} + 7\Delta_i - \Delta_{i+1}}{12}$$

for a general interior point.

Slightly different formulae are applied to the points which do not have two neighbours on both sides, namely

$$d_1 = \frac{25\Delta_1 - 23\Delta_2 + 13\Delta_3 - 3\Delta_4}{12}, \qquad d_2 = \frac{3\Delta_1 + 13\Delta_2 - 5\Delta_3 + \Delta_4}{12},$$

$$d_{N-1} = \frac{\Delta_{N-4} - 5\Delta_{N-3} + 13\Delta_{N-2} + 3\Delta_{N-1}}{12},$$

$$d_N = \frac{-3\Delta_{N-4} + 13\Delta_{N-3} - 23\Delta_{N-2} + 25\Delta_{N-1}}{12}.$$

Similar formulae can be derived for the non-uniform case.

The monotonicity of this cubic interpolant can be enforced by first imposing a necessary condition on the value of the derivatives in the form

$$\begin{aligned} &\operatorname{sign}(d_i) = \operatorname{sign}(\Delta_i) = \operatorname{sign}(d_{i+1}), \quad \Delta_i \neq 0, \\ &d_i = d_{i+1} = 0, \quad \Delta_i = 0, \end{aligned} \tag{3}$$

and then limiting their values[16] in the manner

$$d_i = \operatorname{sign}(d_i) \min(|d_i|, |3\Delta_{i-1}|, |3\Delta_i|). \tag{4}$$

As explained by Rasch and Williamson,[14] (3) and (4) are more than a monotonicity constraint; they are also a form of convexity or positivity constraint in the sense that they control overshoots

on the interval next to local extrema. Being actually non-monotonic in such regions, they prevent oscillations at the edge of flat regions. This can be useful in avoiding clipping of solutions and no new extrema are introduced.

Although there is an inevitable sacrifice of accuracy in the numerical result when monotonicity is sought, the above technique gives satisfactory results for many kinds of problems at a minimum computational cost.

As an illustration, Figures 1 and 2 display some results from the solution of test case 2 with the described monotone Hermite cubic semi-Lagrangian method. They have been computed on two different grids of $N = 13$ and 37 points with $CFL = 1{\cdot}5$. In all cases the continuous line is used as a reference and it represents the solution using $CFL = 1$.

The upper parts of these figures show the temporal variation of the head $H(L, t)$ at the downstream end. The corresponding (spatial) longitudinal head profiles for four dimensionless times (0·125, 1·125, 2·125 and 5·125) are shown in the lower parts of the same figures. They have been arranged so that the thinner the line is, the greater the time it stands for. It can be seen that no oscillations are present in the solutions and that the accuracy increases with the number of points. These results compare very favourably with those published elsewhere.[10]

Unfortunately, this method is not able to give satisfactory results when dealing with open channel flow problems in which discontinuities such as bores or hydraulic jumps may occur.

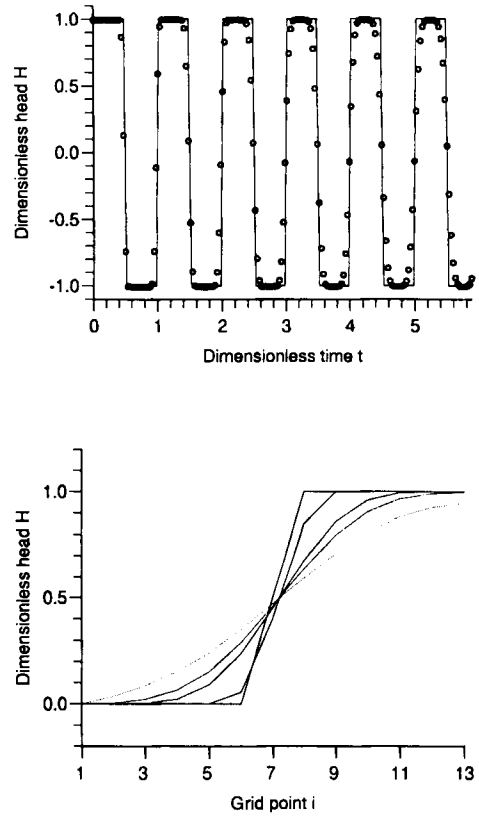

Figure 1. Monotone cubic interpolation. $N = 13$, CFL = 1·5. Upper: downstream head variation with time. Lower: head profiles at various times

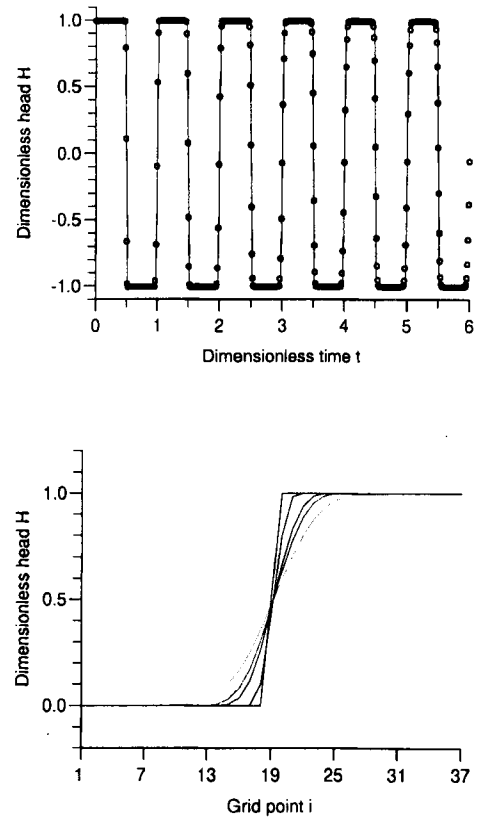

Figure 2. Monotone cubic interpolation. $N = 37$, CFL = 1·5. Upper: downstream head variation with time. Lower: head profiles at various times

This is a consequence of the non-linearity of the equations and hence of the error introduced when trying to solve the ODEs governing the trajectories through the shock.

The numerical results often display a good-looking shape which might be misleading if no exact solution were available. In fact, the solution on both sides of the discontinuity, although monotone, is erroneous and so is the shock speed.

Figures 3 and 4 are examples of this kind of behaviour obtained through the solution of test case 1 for two different values of the initial height ratio $h_L : h_R$. Figure 3 represents the profiles of the water surface 5 s after the dam break in the 5:1 case for two different CFL values. Figure 4 is the equivalent for the 20:1 case after 2·5 s of wave evolution. The available analytic solution appears again as a continuous line.

In the next section a way of improving this situation by rendering the results globally conservative is proposed.

3. RECOVERY OF CONSERVATION

To introduce conservation into the semi-Lagrangian method, reference must be made first to the work of Bermejo and Staniforth.[17] Following an FCT approach,[18,19] they proposed a

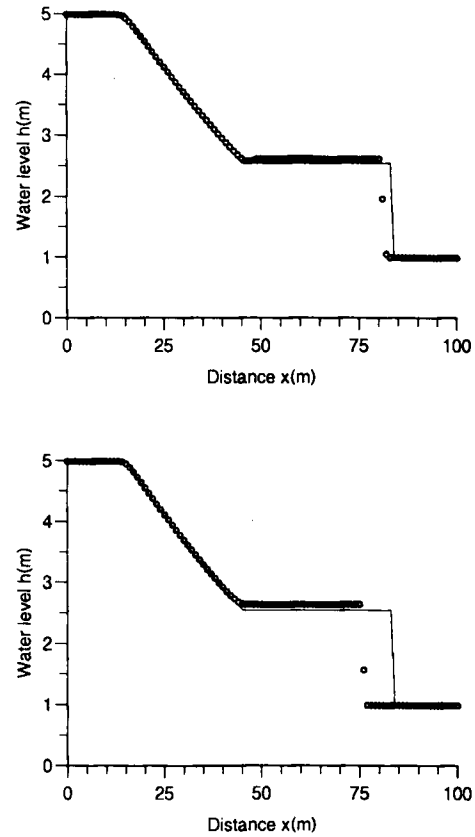

Figure 3. Monotone cubic interpolation. $N = 101$. Dam break problem for a height ratio of 5:1. Upper: CFL = 0·75. Lower: CFL = 1·75

different method of producing schemes able to preserve the shape of the solution near strong gradients whilst maintaining high accuracy in smooth regions, without any special constraint on the interpolation. The basic points underlying this idea can be found in Reference 5. They are now outlined because they will be useful later on in the paper.

A high-order monotone solution U^{M} at the new time level can be defined through a suitable combination of a high-order, possibly oscillatory, solution U^{H} and a low-order, shape-preserving, solution U^{L} as

$$U_i^{M} = \alpha_i U_i^{H} + (1 - \alpha_i) U_i^{L}, \tag{5}$$

with $i = 1, N$ and

$$0 \leq \alpha_i \leq 1. \tag{6}$$

The coefficients $\{\alpha_i\}$ are to be chosen as large as possible whilst maintaining monotonicity. Obviously a trivial solution exists when $\alpha_i = 0$, $i = 1, N$, corresponding to the already monotone U^{L}.

Denoting by U^{n} the solution obtained at the previous time level and by $\{U^{n}, i\}$ the set of

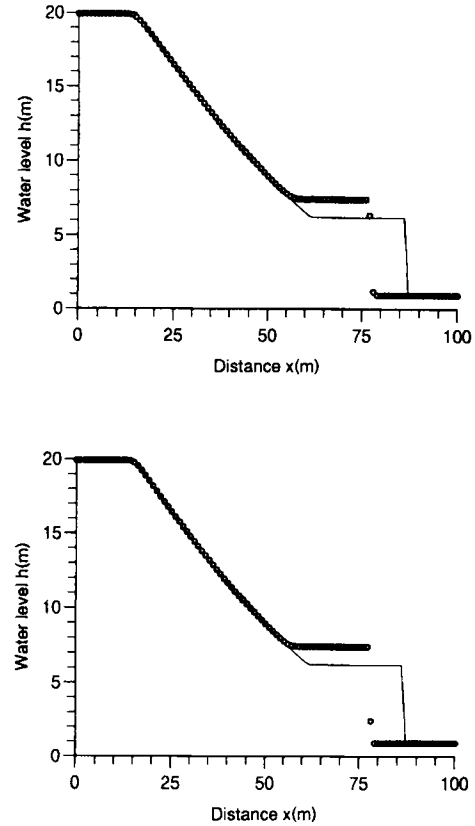

Figure 4. Monotone cubic interpolation. $N = 101$. Dam break problem for a height ratio of 20:1. Upper: CFL = 0·75. Lower: CFL = 1·75

solution values used to interpolate at the foot of the characteristic passing through x_i, the following inequalities provide adequate upper and lower bounds to the coefficients $\{\alpha_i\}$:

$$\min(\{U^n, i\}, U_i^{\mathrm{L}}) \leqslant \alpha_i U_i^{\mathrm{H}} + (1 - \alpha_i) U_i^{\mathrm{L}} \leqslant \max(\{U^n, i\}, U_i^{\mathrm{L}}). \tag{7}$$

Equation (7) differs from the one given by Bermejo and Staniforth[17] only in that the value of the low-order solution has been included in the bounds. When U^{L} is calculated using linear interpolation of the $\{U^n, i\}$ values for pure linear advection problems, the conditions reduce to those of Bermejo and Staniforth. For more general problems including source terms or when solving non-linear systems, the presence of the low-order solution in (7) is necessary to allow new controlled extrema to be generated by the low-order scheme. It is worth noting here that linear interpolation must be used with care when solving non-linear problems and it will be shown to be inadequate for non-linear systems with shocks.

We now proceed to discuss the recovery of conservation and, for the sake of clarity, we shall do it in the scalar case.

3.1 Scalar equations

The set of optimal $\{\alpha_i\}$ satisfying (7) and providing a monotone and accurate solution will be

considered as upper bounds $\{\alpha_i^{\max}\}$ for another choice of $\{\alpha_i\}$ made to produce a monotone, accurate and conservative solution.

In order to find the values

$$0 \leqslant \alpha_i \leqslant \alpha_i^{\max} \tag{8}$$

leading to a conservative solution, the following condition must be imposed on the coefficients:

$$\int U^{\mathrm{M}}(\mathbf{x})\, d\mathbf{x} = \int U^{n}(\mathbf{x})\, d\mathbf{x} = C. \tag{9}$$

Here U^n is the initial solution and C is in general a function of the time and of the boundary conditions. The best that can be done in many cases is, of course, to minimize the difference between the quantities in (9).

The algorithm that is next proposed represents a direct way of obtaining a solution. It must be stressed that the result of seeking a suboptimal $\{\alpha_i\}$ to enforce conservation will sacrifice some accuracy in the results.

Using (7) and defining S_i to be the area associated with the node i (this is just Δx in a regular 1D mesh), (9) can be recast as

$$\sum_i \alpha_i (U_i^{\mathrm{H}} - U_i^{\mathrm{L}}) S_i = C - \sum_i U_i^{\mathrm{L}} S_i = C^*. \tag{10}$$

Let us define

$$\beta_i = (U_i^{\mathrm{H}} - U_i^{\mathrm{L}}) S_i.$$

Then the problem is to maximize the α's subject to the condition

$$\sum_i \alpha_i \beta_i = C^*$$

and the constraints (7) and (8).

For this purpose assume that

$$\sum_i \alpha_i^{\max} \beta_i > C^*. \tag{11}$$

If we had equality in (11), then the monotone and accurate solution would already be conservative. On the other hand, if the inequality were the other way round, some redefinitions ($\beta_i = -\beta_i$, $C^* = -C^*$) could be performed without loss of generality to achieve the satisfaction of (11).

The negative terms of the sum in (11) as well as those equal to zero are supplied with the highest possible coefficient in order to reduce as much as possible the size of the total, i.e.

$$\beta_i \leq 0 \quad \Rightarrow \quad \alpha_i = \alpha_i^{\max}, \quad \textit{if} \; lag(i) = 1.$$

In order to calculate the coefficients for the rest of the terms, an estimate can be made by defining a *surplus* as

$$surplus = C^* - \sum_{i \, flag(i) = 1} \alpha_i \beta_i \tag{12}$$

and an average value of α as

$$\alpha_{AV} = \frac{surplus}{\sum_{if\,lag(i)=0} \beta_i}. \tag{13}$$

If either the *surplus* is negative or all $\beta_i \leqslant 0$, then there is no conservative solution and the best solution as regards conservation is given by the initial set-up of α's. In any other case the average value is compared with α_i^{max} for all the points with $iflag(i) = 0$ and all the values of the coefficients are set equal to the average if it does not exceed the upper bound. This may be expressed as

$$\forall i, \quad if\,lag(i) = 0, \quad \alpha_{AV} < \alpha_i^{max} \quad \Rightarrow \quad \alpha_i = \alpha_{AV}, \quad if\,lag(i) = 1.$$

If, on the contrary, the average value is greater than some of the α_i^{max}, then only those coefficients are fixed and put equal to their maximum value:

$$\forall i, \quad if\,lag(i) = 0, \quad \alpha_{AV} > \alpha_i^{max} \quad \Rightarrow \quad \alpha_i = \alpha_i^{max}, \quad if\,lag(i) = 1.$$

The rest of the points remain with $iflag(i) = 0$.

A new evaluation of (12) and (13) is then performed with a modified number of terms in the sums. The algorithm ends when a value of α_{AV} is found that does not exceed any of the α_i^{max} or when all $iflag(i) = 1$. It works very well for scalar problems, as demonstrated in Reference 20.

This problem can also be solved by linear programming methods. It can be posed in the form of minimizing

$$-\sum_i \alpha_i,$$

where the unknowns are subject to the constraints.

$$0 \leqslant \alpha_i \leqslant \alpha_i^{max}, \qquad \sum_i \alpha_i \beta_i = C^*.$$

3.2. *Systems of equations*

The present section is concerned with the application of the above technique to the 1D shallow water equations. The immediate generalization to systems of equations is to apply it separately to each of the conserved quantities, i.e. the described scalar procedure was used independently to conserve cross-section A and discharge Q in the solution of (1).

In the first attempt, cubic Hermite polynomials and linear interpolation were used respectively as the high-order, U^H, and low-order, U^L, semi-Lagrangian solutions. The unreliability and inadequacy of the linear interpolation as a low-order monotone scheme for this kind of problem were soon realized. Results provided by the use of the linear interpolation semi-Lagrangian scheme as the lower-order solution for two different height ratios are displayed in Figures 5 and 6.

This raised the point of the strong dependence of the quality of the results on the monotonicity of U^L. Moreover, any interpolation scheme which used the information coming through the characteristics was likely to demonstrate the same behaviour across a shock.

In order to overcome this difficulty without losing the advantages offered by semi-Lagrangian schemes, a generalization of the first-order Roe method,[21] modified to allow large time steps, was explored[22,23] first in the scalar non-linear case and then adapted to systems of equations. The difference scheme is extended by explicitly handling the interactions of the solutions to the Riemann problems at each interface. It becomes stable for $CFL > 1$ and provides an accurate

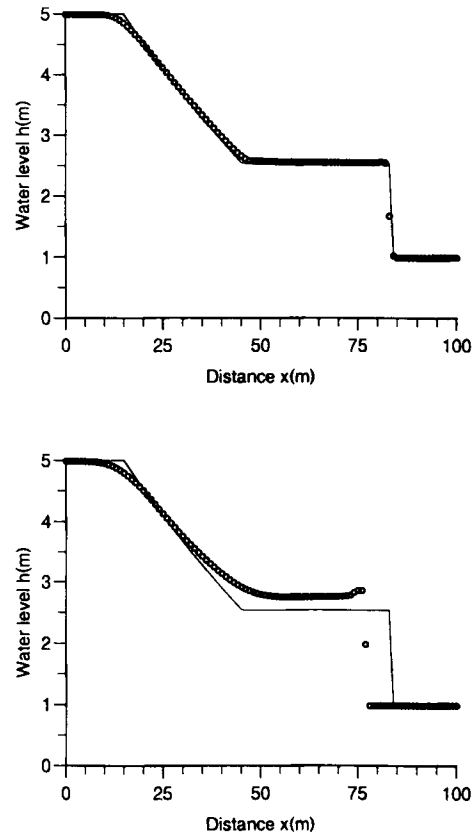

Figure 5. Recovery of conservation: cubic and linear interpolation. $N = 101$. Dam break problem for a height ratio of 5:1. Upper: CFL = 0·75. Lower: CFL = 1·75

and correct solution of shocks. In the case of a linear problem it reduces to a linear interpolation scheme.

Provided that Roe's linearization is used to decouple the system (1), here expressed as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = 0, \qquad A = \frac{\partial \mathbf{F}}{\partial \mathbf{U}},$$

an approximate matrix A^* can be built whose eigenvalues (λ^1, λ^2) and eigenvectors $(\mathbf{e}^1, \mathbf{e}^2)$ satisfy

$$\delta \mathbf{U}_{i+1/2} = \mathbf{U}_{i+1} - \mathbf{U}_i = \sum_k a^k \mathbf{e}^k, \qquad \delta \mathbf{F}_{i+1/2} = \mathbf{F}_{i+1} - \mathbf{F}_i = \sum_k \lambda^k a^k \mathbf{e}^k.$$

Expressions for λ^k, a^k and $\mathbf{e}^k$ can be found, for instance, in Reference 21.

The basic idea is to calculate $\delta \mathbf{U}$ at every interface and update the different k-waves according to the sign of their celerities and the values of the local CFL numbers. As an example, if at $i+\frac{1}{2}$, for $k = 1$, $\lambda^1_{i+1/2} > 0$, then each δU_j, $j = 1, 2$, will affect the corresponding variable U_j so that

$$a^1 e^1_j \text{ is added to } i+1, \ldots, i+\mu^1, \qquad (v^1 - \mu^1) a^1 e^1_j \text{ is added to } i+\mu^1+1,$$

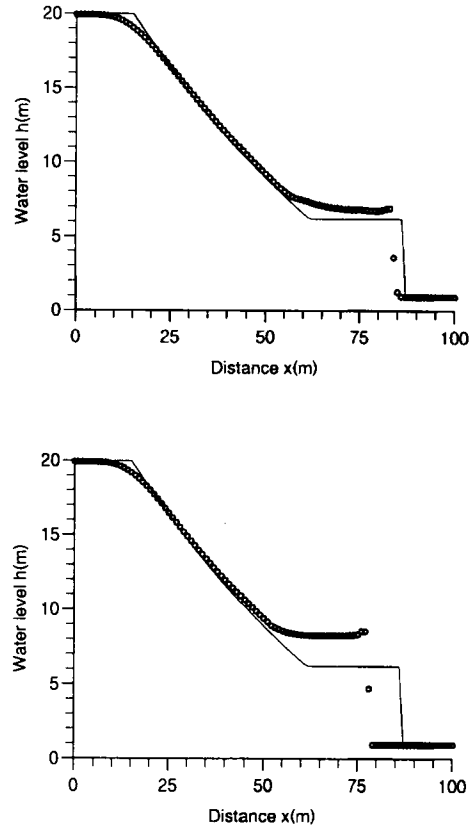

Figure 6. Recovery of conservation: cubic and linear interpolation. $N = 101$. Dam break problem for a height ratio of 20:1. Upper: CFL = 0·75. Lower CFL = 1·75

where

$$v^1 = \frac{\Delta t}{\Delta x}(\lambda^1)_{i+1/2}, \qquad \mu^1 = \text{int}(v^1).$$

In order to avoid the appearance of non-physical shocks, rarefaction waves are split[24] and propagated in both directions.

The improvement achieved with the use of this technique as the low-order solution for the dam break test case can be observed in Figures 7 and 8.

The application of the same procedure to the water hammer problem brought to light new difficulties arising from the strategy of enforcing conservation separately in both variables. The consequence of having a different set of coefficients $\{\alpha_i\}$ for each variable is that a slight phase shift between them appears which is more noticeable in this case test, because there is an interaction with the boundaries which is not present in the dam break problem. Examples of the distortion generated by this approach can be observed in Figure 9.

Thus the next question is: can a common set of α's be used for both variables whilst still conserving both variables? As a partial answer to this question, an extension of the recovery-of-conservation algorithm will be proposed in the next section and the numerical results of its implementation will be presented.

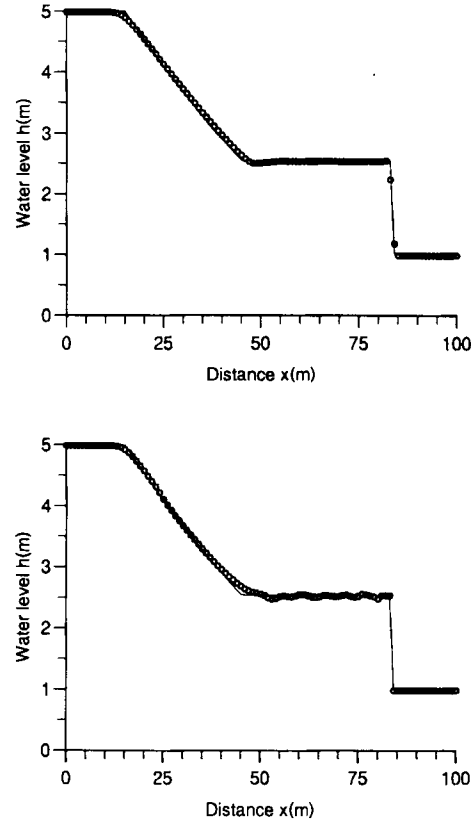

Figure 7. Recovery of conservation: cubic interpolation and Roe's scheme. $N = 101$. Dam break problem for a height ratio of 5:1. Upper: CFL = 0·75. Lower: CFL = 1·75

4. IN-PHASE CONSERVATION

There is some freedom in the way of solving the problem of finding common coefficients for both variables. The simplest solution is to enforce conservation in one of them, say the variable $F1$, and applying the set of coefficients $\{\alpha_i^{(1)}\}$ to the variable $F2$. This obviously removes the phase shift and ensures conservation in $F1$. Conservation in $F2$, although improved, is nevertheless not enforced in this way.

A better way is to require conservation in one variable whilst trying to minimize the error in conservation in the other, using a common set of coefficients which will be called $\{\alpha_i^c\}$ from now on. To do this, the scalar mechanism described in Section 3 is applied to one of the variables, $F1$, up to the moment at which, having fixed some of the coefficients $\alpha_i^{(1)}$, a suitable value α_{AV} for the rest of the points is found which ensures $F1$ conservation. From that moment the algorithm differs mainly in that the variable $F2$ will also be involved in the calculation of the optimal α^c's.

To begin with, the new coefficients are defined as

$$\alpha_i^c = \begin{cases} \alpha_i^{(1)} & \text{if } iflag(i) = 1, \\ \alpha_{AV} & \text{if } iflag(i) = 0, \end{cases}$$

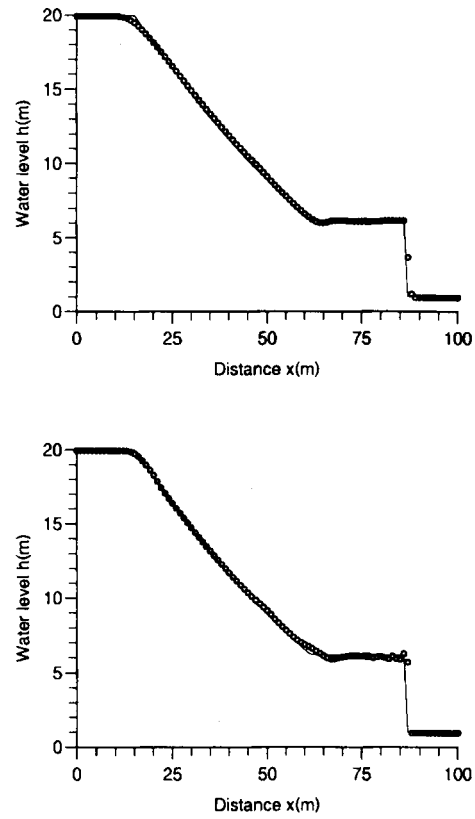

Figure 8. Recovery of conservation: cubic interpolation and Roe's scheme. $N = 101$. Dam break problem for a height ratio of 20:1. Upper: CFL = 0·75. Lower: CFL = 1·75

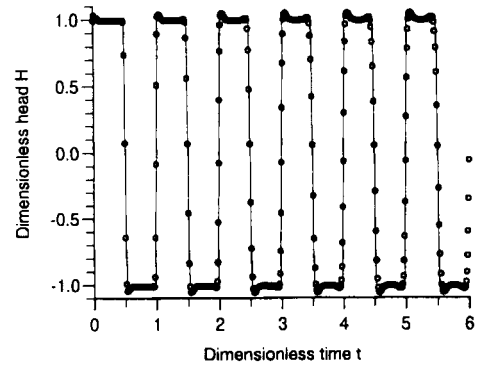

Figure 9. Recovery of conservation: cubic interpolation and Roe's scheme. Water hammer problem. $N = 37$, CFL = 1·5

so that from conservation in $F1$ the following equality holds:

$$\sum_i \alpha_i^c \beta_i^{(1)} + \Delta x \sum_i F_i^{(1)L} = C1. \tag{14}$$

However, now we have in general

$$\sum_i \alpha_i^c \beta_i^{(2)} + \Delta x \sum_i F_i^{(2)L} \neq C2 \tag{15}$$

and our problem is to approach as closely as possible the equality in (15) without violating (14).

In the first sum in (15) the terms corresponding to $iflag(i) = 1$ are fixed. Those with $iflag(i) = 0$ will allow us to carry out some adjustments. They will be grouped so that (15) is expressed as

$$S_0 + S_1 + S_L \neq C2,$$

where

$$S_0 = \sum_{iflag(i)=0} \alpha_i^c \beta_i^{(2)}, \qquad S_1 = \sum_{iflag(i)=1} \alpha_i^c \beta_i^{(2)}, \qquad S_L = \Delta x \sum_i F_i^{(2)L},$$

or even more briefly as

$$S_0 \neq target,$$

where $target = C2 - S_1 - S_L$.

Suppose that

$$S_0 > target \tag{16}$$

(otherwise (16) could always be arranged by redefining $C2$, $\{\beta_i^{(2)}\}$ and S_L). Then the objective is to minimize S_0 subject to (16), (14) and the conditions (8) imposed by the upper and lower limits on the values of the coefficients.

For that purpose the set of values $\{\beta_i^{(2)}\}$ with $iflag(i) = 0$ can be ordered so that the smallest (corresponding to $i = is$) and largest (corresponding to $i = ib$) values can be selected. We would like to reduce α_{ib}^c as much as possible in order to diminish the weight of the biggest term in S_0, taking into account that

$$\alpha_{is}^c \beta_{is}^{(1)} + \alpha_{ib}^c \beta_{ib}^{(1)} = T^{(1)} \tag{17}$$

must hold.

If it happens that

$$\alpha_{is}^{max} \beta_{is}^{(1)} > T^{(1)},$$

then α_{ib}^c can be completely removed from (17), so that

$$\alpha_{ib}^c = 0, \quad iflag(ib) = 1,$$

$$\alpha_{is}^c = \frac{T^{(1)}}{\beta_{is}^{(1)}}, \quad iflag(is) = 0.$$

Otherwise α_{is}^c can be supplied with the largest possible value:

$$\alpha_{is}^c = \alpha_{is}^{max}, \quad iflag(is) = 1,$$

$$\alpha_{ib}^c = \frac{T^{(1)} - \alpha_{is}^{max} \beta_{is}^{(1)}}{\beta_{ib}^{(1)}}, \quad iflag(ib) = 0.$$

In both cases α_{is}^c is reduced and α_{ib}^c increased. With a different number of points at which $iflag(i) = 0$, the sums S_0 and S_1 are re-evaluated to see whether the inequality (16) still maintains its sign. If so, the process is repeated with another couple until either all $iflag(i) = 1$ or (16) changes its sign.

If a certain couple of values (is, ib) produces

$$S_0 < target,$$

it means that an adequate combination of $\beta_{is}^{(2)}$ and $\beta_{ib}^{(2)}$ will provide conservation in both variables. Suitable coefficients can be calculated from the two conditions they must fulfil, namely

$$\alpha_{is}^c \beta_{is}^{(1)} + \alpha_{ib}^c \beta_{ib}^{(1)} = T^{(1)}, \qquad \alpha_{is}^c \beta_{is}^{(2)} + \alpha_{ib}^c \beta_{ib}^{(2)} = T^{(2)},$$

where

$$T^{(2)} = C2 - S_L - \sum_{i \neq is, ib} \alpha_i^c \beta_i^{(2)}$$

always constrained by their limiting values. Testing must be continually performed to ensure that the α^cs do not violate the constraints.

This problem is again a good candidate for solution by a simple linear programming procedure. In this case it would consist of minimizing the sum over the points with $iflag(i) = 0$, i.e.

$$-\sum_i \alpha_i^c,$$

with the α_i^c subject to the constraints

$$0 \leqslant \alpha_i^c \leqslant \alpha_i^{\max}, \qquad \sum_i \alpha_i^c \beta_i^{(1)} = RS1, \qquad \sum_i \alpha_i^c \beta_i^{(2)} \geqslant target,$$

where $RS1 = C1 - \Delta x \Sigma_i F_i^{(1)L}$.

The results of the performance of the Hermite cubic polynomial interpolation semi-Lagrangian scheme supplied with an in-phase conservation recovery using Roe's scheme as the low-order solution are illustrated in Figures 10–12. The improvement introduced by the in-phase algorithm over the previous way of recovering conservation for the water hammer test case is evident from Figures 10 and 11. We remark that the results are the same whatever variable is used as $F1$ to enforce conservation, because this problem is symmetric in both variables H and V.

In order to summarize as well as quantitatively compare the methods considered in this paper, the measures of CPU time consumed on a SUN SPARC2 workstation and accuracy achieved by them in two particular cases have been tabulated. The accuracy has been estimated by means of an L_2-norm defined as

$$L_2(u) = \sqrt{\left(\sum_i \Delta x (u_i^e - u_i^n)^2\right)},$$

where u^e and u^n represent the exact and numerical solutions respectively.

First, the results for the linear water hammer problem on four different grids using $CFL = 0{\cdot}5$ are contained in Tables I and II. The abbreviations in the first column stand for: HP, Hermite polynomials; MHP, monotone Hermite polynomials; IRC, independent recovery of conservation; IRCLP, linear programming of the independent recovery of conservation; IPRC, in-phase recovery of conservation; IPRCLP, linear programming of the in-phase recovery of conservation.

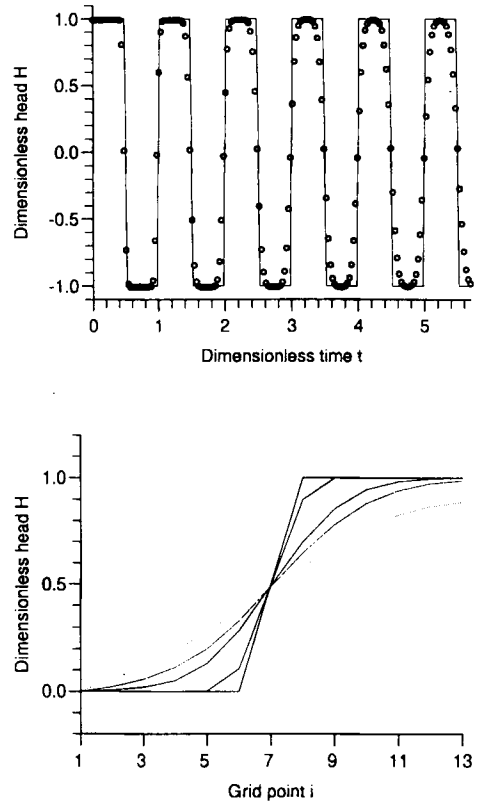

Figure 10. In-phase recovery of conservation: cubic interpolation and Roe's scheme. Water hammer problem. $N = 13$, CFL = 1·5. Upper: downstream head variation with time. Lower: head profiles at various times

Table I. CPU time used by the different methods in the linear problem

	$N = 12$	$N = 24$	$N = 48$	$N = 96$
HP	3·1	11·7	44·7	175·3
MHP	3·2	12·4	48·7	193·8
IRC	5·3	19·3	73·8	288·7
IRCLP	8·7	24·5	88·6	354·6
IPRC	6·2	21·9	83·5	316·9
IPRCLP	7·8	21·4	68·8	229·0

The equivalent tables for the solution of the non-linear dam break problem in the case of an initial height ratio of 5:1 using $CFL = 0{\cdot}8$ are presented next (Tables III–IV).

The basic conservation recovery algorithm represents a significant saving over the linear programming solution. Compared with the MHP scheme, which they both use as starting point, our algorithm is twice as fast for the linear problem (see Table I). For the non-linear problem the savings are less spectacular, but the linear programming technique is still roughly 50 per cent faster.

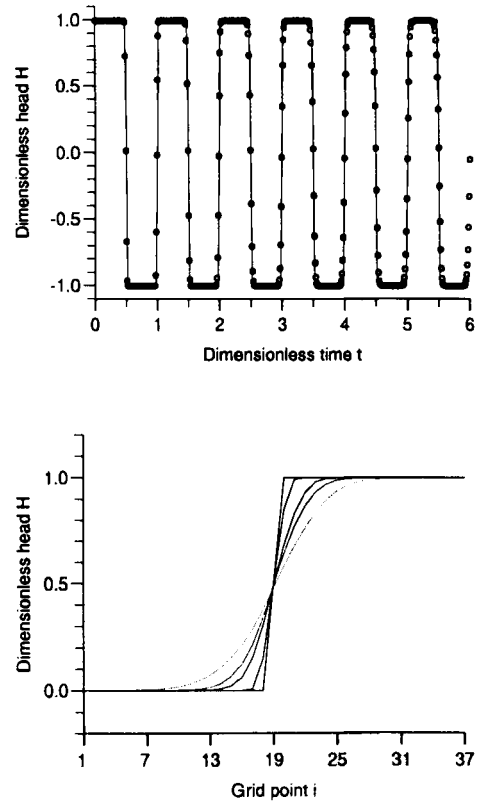

Figure 11. In-phase recovery of conservation: cubic interpolation and Roe's scheme. Water hammer problem. $N = 37$, CFL = 1·5. Upper: downstream head variation with time. Lower: head profiles at various times

Table II. L_2-error in H in the linear problem

	$N = 12$	$N = 24$	$N = 48$	$N = 96$
HP	0·629	0·384	0·332	0·269
MHP	0·795	0·489	0·387	0·304
IRC	0·684	0·535	0·415	0·322
IRCLP	0·718	0·556	0·426	0·327
IPRC	0·857	0·629	0·461	0·348
IPRCLP	0·832	0·625	0·456	0·365

With the in-phase recovery the linear programming version is generally but not always faster. Since the in-phase recovery algorithm is also more difficult to programme, we feel that the linear programming problem is the one to solve in this case when linear programming routines are available. Otherwise the algorithm presented here is a useful alternative.

In Reference 10 it was shown that the basic semi-Lagrangian method with a cubic interpolant was better than other schemes used in the hydraulics literature for the water hammer problem. We see in Table II that some accurcy is lost in gaining conservation, although the loss is

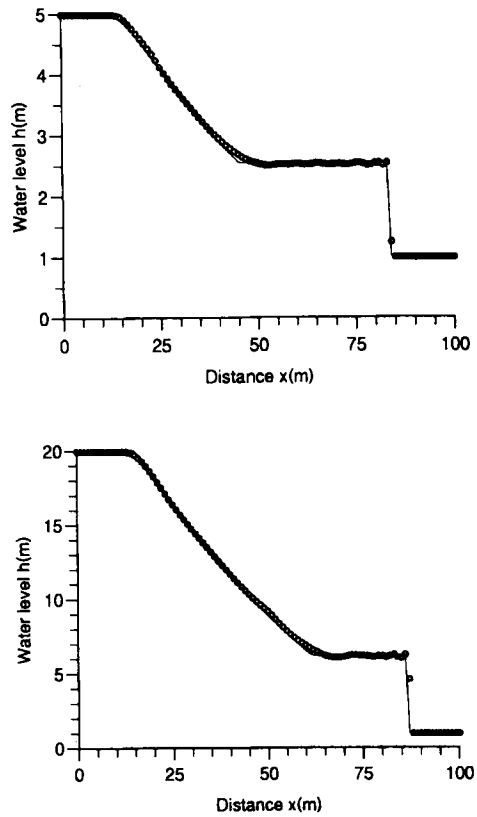

Figure 12. In-phase recovery of conservation: cubic interpolation and Roe's scheme. $N = 101$. Dam break problem. CFL = 1·75. Upper: height ratio of 5:1. Lower: height ratio of 20:1

Table III. CPU time used by the different methods in the non-linear problem

	$N = 50$	$N = 100$	$N = 200$	$N = 400$	$N = 800$
MHP	1·1	5·6	30·6	190·7	1051·3
IRC	1·6	7·7	39·0	226·4	1467·7
IRCLP	2·1	9·3	44·1	244·0	1549·0
IPRC	1·8	8·4	41·5	229·1	1469·1
IPRCLP	2·0	8·9	40·9	220·8	1446·7

somewhat smaller than the errors incurred in achieving monotonicity. It is worth mentioning here that the difference in errors for the IRC and IRCLP methods are caused by the fact that the solution to the linear programming problem is not unique. They could, if desired, be forced to produce the same answers.

The in-phase conservation has larger errors because smaller values of α^{max} must be taken, the method then using less of the higher-order scheme.

The reason for using these recovery techniques becomes more evident in Tables IV and V,

Table IV. L_2-error in A in the non-linear problem

	$N = 50$	$N = 100$	$N = 200$	$N = 400$	$N = 800$
MHP	1·09	1·909	1·937	1·961	1·974
IRC	1·069	0·438	0·295	0·328	0·125
IRCLP	1·220	0·409	0·383	0·454	0·230
IPRC	0·953	0·594	0·500	0·448	0·309
IPRCLP	0·940	0·699	0·623	0·368	0·267

where we see that the monotone version of the semi-Lagrangian method is not converging to the correct answer. The various recovery algorithms cure this problem.

5. CONCLUSIONS

Monotone Hermite cubic polynomials with derivatives calculated explicitly from the neighbouring points seem a very efficient method of interpolation in the context of semi-Lagrangian schemes.

A new recovery-of-conservation procedure has been proposed to render the scheme much more suited to hydraulic problems with shocks. It is based on an FCT approach and relies on the adequate choice of a set of coefficients combining a high- and a low-order method. Two hydraulic problems have been used as test cases, water hammer and dam break. In the latter case special care is needed with the low-order solution and we have shown the advantages of using Leveque's large time step version of Roe's scheme for this purpose.

In this paper we have been concerned with 1D shallow water equations and so have made use of Riemann invariants. Most of the concepts presented here extend, nevertheless, to higher dimensions and to large systems. The semi-Lagrangian approach and the basic independent recovery algorithms work equally well in higher dimensions.[22,25] They can clearly be applied to arbitrarily large systems. The in-phase conservation, in principle, could also be extended, but there is a real danger of running out of degrees of freedom before conservation in all required variables is achieved.

Although the result is not a shock-capturing technique, it is able to correctly reproduce the discontinuities of interest in common engineering pipe calculations, as we have shown, also allowing the use of time steps not restricted by the usual CFL stability conditions.

ACKNOWLEDGMENTS

We would like to thank Dr. Mike Baines for many useful and interesting discussions during the course of this work. The first author was funded by the Spanish Ministerio de Educacion y Ciencia and the second by the U.K. Science and Engineering Research Council.

REFERENCES

1. A. Staniforth and J. Cote, 'Semi-Lagrangian integration schemes for atmospheric models—a review', *Mon. Weather Rev.*, **119**, 2206–2223 (1991).
2. R. Courant, E. Isaacson and M. Rees, 'On the solution of nonlinear hyperbolic differential equations by finite differences', *Commun. Pure Appl. Math.*, **5**, 243–255 (1952).
3. D. C. Wiggert and M. J. Sundquist, 'Fixed-grid characteristics for pipeline transients', *J. Hydraul. Div. ASCE*, **103**, 1403–1415 (1978).
4. N. Katopodes and D. R. Schamber, 'Applicability for dam-break flood wave models', *J. Hydraul. Eng.*, **109**, (1983).

5. A. Priestley, 'A quasi-Riemannian method for the solution of one-dimensional shallow water flow', *J. Comput. Phys.*, **106,** 139–147 (1993).
6. K. W. Morton, A. Priestley and E. Suli, 'Stability of the Lagrange–Galerkin method with non-exact integration', *Math. Modell. Numer. Anal.*, **22,** 625–653 (1988).
7. R. Bermejo, 'On the equivalence of semi-Lagrangian schemes and particle-in-cell finite element methods', *Mon. Weather Rev.*, **118,** 979–987 (1990).
8. R. K. Smolarkiewicz and G. A. Grell, 'A class of monotone interpolation schemes', *J. Comput. Phys.*, **101,** 431–440 (1992).
9. M. Gomez Valentin, 'A hydraulic numerical model for the analysis of unsteady flow in hydroelectric canals', *Proc. XXIV IAHR Congr.*, Madrid, 1991.
10. I. A. Sibetheros and E. R. Holley, 'Spline interpolations for water hammer analysis', *J. Hydraul. Eng.*, **117,** 1332–1351 (1991).
11. J. D. Lambert, *Computational Methods in Ordinary Differential Equations*, Wiley, New York, 1973.
12. G. A. Schohl and F. M. Holly Jr., 'Cubic-spline interpolation in Lagrangian advection computation', *J. Hydraul. Eng.*, **117,** 248–253 (1991).
13. F. M. Holly Jr. and A. Preissmann, 'Accurate calculation of transport in two dimensions', *J. Hydraul. Div., ASCE*, **103,** 1259–1277 (1977).
14. P. J. Rasch and D. L. Williamson, 'On Shape-preserving Interpolation and Semi-Lagrangian Transport', *SIAM J. Sci. Stat. Comput.*, **11** (4), 1990.
15. J. M. Hyman, 'Accurate monotonicity preserving cubic interpolations', *SIAM J. Sci. Stat. Comput.*, **4,** 645–654 (1983).
16. C. De Boor and B. Swartz, 'Piecewise monotone interpolation', *J. Approx. Theory*, **21,** 411–416 (1977).
17. R. Bermejo and A. Staniforth, 'The conversion of semi-Lagrangian advection schemes to quasi-monotone schemes', *Mon. Weather Rev.*, **120,** 2622–2632 (1992).
18. J. P. Boris and D. L. Book, 'Flux corrected transport, I, SHASTA, a fluid transport algorithm that works', *J. Comput. Phys.*, **11,** 38–69 (1973).
19. S. T. Zalesak, 'Fully multidimensional flux-corrected transport algorithms for fluids', *J. Comput. Phys.*, **31,** 335–362 (1979).
20. A. Priestley, 'A quasi-conservative version of the semi-Lagrangian advection scheme', *Mon. Weather Rev.*, **121,** 621–629 (1993).
21. P. L. Roe, 'Approximate Riemann solvers, parameter vectors and difference schemes', *J. Comput. Phys.*, **43,** 357–372 (1981).
22. R. J. Leveque, 'A large time step generalization of Godunov's method for systems of conservation laws', *SIAM J. Numer. Anal.*, **33,** 1051–1073 (1971).
23. R. J. Leveque, 'Large time step shock-capturing techniques for scalar conservation laws', *Numerical Analysis Project, Manuscript NA-81-13*, Stanford University, 1981.
24. A. Harten and P. Hyman, 'Self adjusting grid methods for one-dimensional hyperbolic conservation laws', *J. Comput. Phys.*, **50,** 235–269 (1982).
25. S. Gravel and A. Staniforth, 'A mass conserving semi-Lagrangian scheme for the shallow-water equations', *Mon. Weather Rev.*, (to appear).