INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2000; **33**: 583–608

The anti-dissipative, non-monotone behavior of Petrov–Galerkin upwinding

Scott F. Bradford[a] and Nikolaos D. Katopodes[b,*]

[a] *Naval Research Laboratory, Washington, DC, U.S.A.*
[b] *Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, U.S.A.*

SUMMARY

The Petrov–Galerkin method has been developed with the primary goal of damping spurious oscillations near discontinuities in advection dominated flows. For time-dependent problems, the typical Petrov–Galerkin method is based on the minimization of the dispersion error and the simultaneous selective addition of dissipation. This optimal design helps to dampen the oscillations prevalent near discontinuities in standard Bubnov–Galerkin solutions. However, it is demonstrated that when the Courant number is less than 1, the Petrov–Galerkin method actually amplifies undershoots at the base of discontinuities. This is shown in an heuristic manner, and is demonstrated with numerical experiments with the scalar advection and Richards' equations. A discussion of monotonicity preservation as a design criterion, as opposed to phase or amplitude error minimization, is also presented. The Petrov–Galerkin method is further linked to the high-resolution, total variation diminishing (TVD) finite volume method in order to obtain a monotonicity preserving Petrov–Galerkin method. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: non-monotone; Petrov–Galerkin; upwinding method

1. INTRODUCTION

The development of spurious oscillations near discontinuities in advection-dominated flows has long been a problem in numerical modeling. Often, some form of upwinding is utilized to dampen the over- and undershoots. In particular, the Petrov–Galerkin method has often been used when attempting to find a finite element solution of the governing equations. Such a technique has been applied to the advection–diffusion equation by Hughes and Brooks [1], and to the shallow-water equations by Katopodes [2].

For a time-dependent problem, Raymond and Garder [4] specifically designed a Petrov–Galerkin method to be slightly more dissipative than the standard Bubnov–Galerkin method,

* Correspondence to: Department of Civil and Environmental Engineering, The University of Michigan, College of Engineering, 116 EWRE Building, 1351 Beal Avenue, Ann Arbor, MI 48109-2125, U.S.A.

Received 4 March 1997
Revised 13 August 1999

while possessing greater phase accuracy. This selective dissipation allows for the damping of the high frequency oscillatory waves found near discontinuities, while not affecting the low frequency long waves associated with the solution of shallow-water equations, for example.

However, there are instances when the Petrov–Galerkin scheme can actually be *less* dissipative than the Bubnov–Galerkin method at certain locations in the domain. Specifically, this can occur when the Courant number is less than 1. For such cases, both the Petrov and Bubnov–Galerkin solutions tend to develop undershoots at the base of a discontinuity. However, the Petrov–Galerkin perturbation of the weighting function tends to exacerbate such oscillations, sometimes making their amplitude more than twice as large as in the Bubnov–Galerkin solution. This fact is not revealed by the Fourier analysis performed on the semi-discrete equations by Raymond and Garder [4] or in the analysis of the fully discrete case for the kinematic wave equation presented by Katopodes [3], in which it is shown that the Petrov–Galerkin scheme is at least as dissipative as the Bubnov–Galerkin method for all wave frequencies and all Courant numbers.

Alternatively, the undershoots could be avoided altogether by designing a scheme that preserves the monotonicity of the solution. This has been the focus of the high resolution, total variation diminishing (TVD), finite volume method that has been widely applied in the field of unsteady gas dynamics. More recently, Bøe [5] derived a monotone Petrov–Galerkin method for the quasi-linear, advection–diffusion equation utilizing these ideas. It will be demonstrated that a Petrov–Galerkin method can be constructed such that it is equivalent to the TVD method. Such a scheme can then be utilized to find monotonicity preserving solutions to advection-dominated problems.

2. SCALAR ADVECTION

The anti-dissipative nature of the Petrov–Galerkin method is first illustrated for the case of scalar advection in one space dimension. The scalar conservation equation can be written as

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{1}$$

where u is a scalar quantity, a is a constant advection velocity, t denotes time, and x is the spatial co-ordinate.

The finite element discretization is obtained by multiplying Equation (1) by a weighting function, W_i, and integrating over the domain with the assumption that u varies linearly over each element, i.e.

$$u(x, t) \approx \sum_{j=1}^{2} N_j(x)u_j(t) \tag{2}$$

where j denotes the node of the element and N_j are the standard chapeau trial functions.

For the Bubnov–Galerkin scheme, $W_i = N_i$. However, for the Petrov–Galerkin method, the weighting function is chosen as

$$W_i = N_i + p\frac{\partial N_i}{\partial x} \tag{3}$$

where p is an undetermined parameter.

In addition, if Crank–Nicolson time stepping is utilized, the assembly of two consecutive elements of identical size yields the following finite difference equation at an interior global node j [3]:

$$\frac{1}{6}\dot{u}_{j-1} + \frac{2}{3}\dot{u}_j + \frac{1}{6}\dot{u}_{j+1} - p\left(\frac{\dot{u}_{j+1} - \dot{u}_{j-1}}{2\Delta x}\right) + \frac{a}{2}\left(\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x} + \frac{u_{j+1}^{n} - u_{j-1}^{n}}{2\Delta x}\right)$$
$$- \frac{ap}{2}\left(\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} + \frac{u_{j+1}^{n} - 2u_j^{n} + u_{j-1}^{n}}{\Delta x^2}\right) = 0 \tag{4}$$

where

$$\dot{u}_j = \frac{u_j^{n+1} - u_j^n}{\Delta t} \tag{5}$$

and n denotes the current time level. From Equation (4), it can be seen that this is merely the Galerkin discretization of the following equation:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = p\frac{\partial^2 u}{\partial x\,\partial t} + ap\frac{\partial^2 u}{\partial x^2} \tag{6}$$

The Fourier analysis performed by Raymond and Garder [4] showed that the optimal value of p is such that

$$p \equiv r\Delta x = \frac{\Delta x}{\sqrt{15}} \tag{7}$$

This particular choice reduces the phase error of the scheme from $O(\omega^4)$ for the Bubnov–Galerkin method to $O(\omega^6)$, where ω is the wave frequency. The sign of p is chosen to match the sign of a, which insures that the scheme is dissipative. This can also be seen by examination of the right-hand side of Equation (6) in which the term ap multiplies a diffusion term. Thus, in order to remain diffusive, this coefficient must always remain positive. In fact, this choice of p introduces a dissipative error of $O(\omega^4)$ as opposed to the error of $O(\omega^6)$ associated with the Bubnov–Galerkin method.

Thus, the second term on the right-hand side of Equation (6) is an artificial diffusion term, which helps dampen the high frequency waves associated with discontinuities. The first term on the right-hand side of Equation (6) provides the improved phase accuracy of the Petrov–Galerkin method. However, under certain circumstances and at specific locations within the domain, this term can dominate the artificial diffusion term, creating an anti-dissipative scheme.

This is more clearly illustrated if Equation (6) is rewritten in terms of the material derivative as follows:

$$\frac{\mathrm{D}u}{\mathrm{D}t} = Q \tag{8}$$

where

$$Q = p\frac{\partial^2 u}{\partial x\,\partial t} + ap\frac{\partial^2 u}{\partial x^2} \tag{9}$$

acts as a source or sink of u, depending upon the sign of Q. If $Q > 0$ then a parcel moving with speed a will gain u, while if $Q < 0$, then there is a loss of u.

In order to quantify this theory, consider the scenario illustrated in Figure 1, where at $t = 0$ there exists a discontinuous pulse of u with amplitude $u_{\mathrm{L}} - u_{\mathrm{R}} > 0$ (denoted by the solid line). It is assumed that $a > 0$ and $a\Delta t \leq \Delta x$. At $t = \Delta t$, the pulse is assumed to travel a distance $a\Delta t$ as shown by the dotted line. This is an approximation that is only valid in the limit $Q \to 0$.

With this assumption and the aid of Equation (4), the two terms on the right of Equation (9) are approximated as

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{\alpha \Delta u}{2\Delta x^2} \tag{10}$$

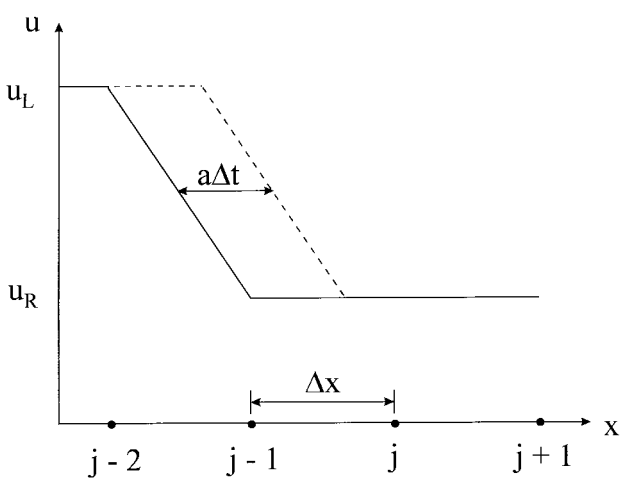

Figure 1. Advection of a discontinuous pulse of a scalar, u towards node j. Solid line denotes initial condition of u (at $t = 0$) and dashed line denotes the state of u at $t = \Delta t$.

$$\frac{\partial^2 u}{\partial x\,\partial t} \approx -\frac{\alpha \Delta u}{2\Delta x\,\Delta t} \tag{11}$$

where $\alpha = a\Delta t/\Delta x$ is the Courant number and $\Delta u = u_L - u_R$. Q is then estimated at node j as

$$Q_j \approx \frac{r\alpha \Delta u}{2\Delta t}(\alpha - 1) \tag{12}$$

which can be generalized for the case when $a < 0$ and $\Delta u < 0$, as

$$Q_j \approx \frac{|r\alpha \Delta u|}{2\Delta t}(|\alpha| - 1) \tag{13}$$

which is clearly always negative for $|\alpha| < 1$. Therefore, when $|\alpha| < 1$, Q_j becomes a sink for u and undershooting should be observed at the base of the discontinuity. However, as $\alpha \to 1$, it is expected that the oscillations would diminish. For $\alpha > 1$, Q_j becomes a source for u and the sharp wave should be smeared at its base.

The previous example illustrates what happens at the base of a discontinuity as it reaches a given node j. In order to examine what happens as the discontinuity passes through node j, two additional examples are considered, as illustrated by the alternative initial conditions presented in Figures 2 and 3. For the case in Figure 2, the discontinuity is passing through node j and Q_j is approximated as

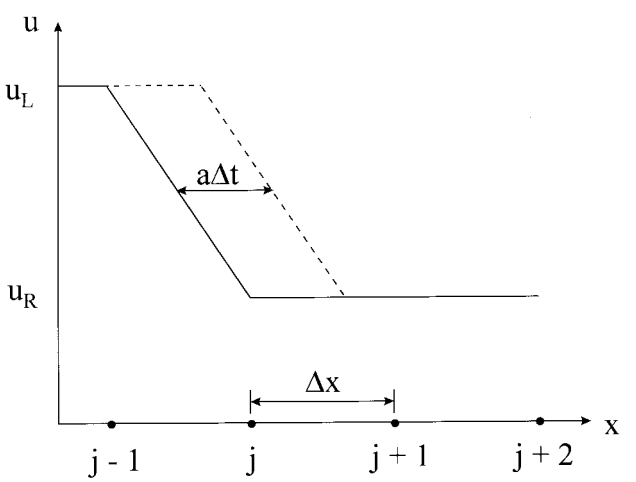

Figure 2. Advection of a discontinuous pulse of a scalar u through node j. Solid line denotes initial condition of u (at $t = 0$) and dashed line denotes the state of u at $t = \Delta t$.

$$Q_j \approx \frac{|r\alpha\Delta u|}{\Delta t}(1-|\alpha|) \tag{14}$$

which is always positive for $|\alpha|<1$. Therefore, there should be a gain of u at node j in this instance. For case in Figure 3, the discontinuity is passing by node j and Q_j is approximated as

$$Q_j \approx \frac{|r\alpha\Delta u|}{2\Delta t}(|\alpha|-1) \tag{15}$$

which is always negative for $|\alpha|<1$, and therefore there should be a decrease of u at node j as in the first case. This alternating pattern of overshooting and undershooting at consecutive nodes spanning a discontinuity is a notorious problem associated with finite element solutions.

The previous analysis illustrates how the Petrov–Galerkin method behaved when simulating the advection of a scalar discontinuity. However, an approximate solution at the $n+1$ time level was utilized in the estimation of Q_j. In order to illustrate the undershooting problem without this limiting assumption, consider the following simple case illustrated in Figure 4. Assuming the initial condition that $u(x, t=0)=0$ for $0 \le x \le L$, where L is the length of the domain, and given appropriate boundary conditions, then the unknowns at $t=\Delta t$ can be computed utilizing the previously presented finite element discretization.

A typical treatment for inputing a scalar at the left boundary is a Dirichlet condition, i.e.

$$u(x=0, t)=U \tag{16}$$

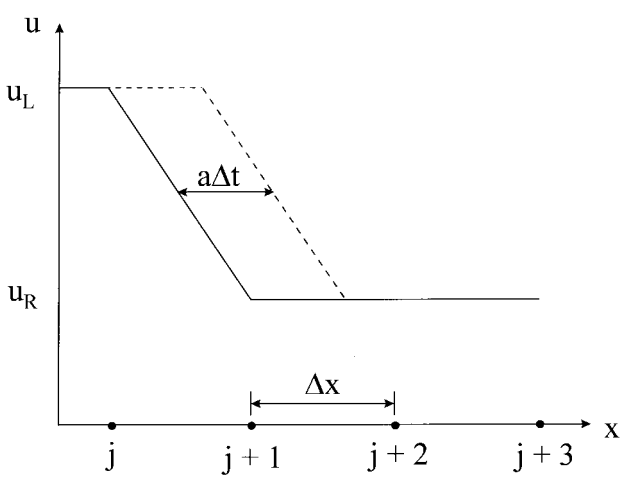

Figure 3. Advection of a discontinuous pulse of a scalar u past node j. Solid line denotes initial condition of u (at $t=0$) and dashed line denotes the state of u at $t=\Delta t$.

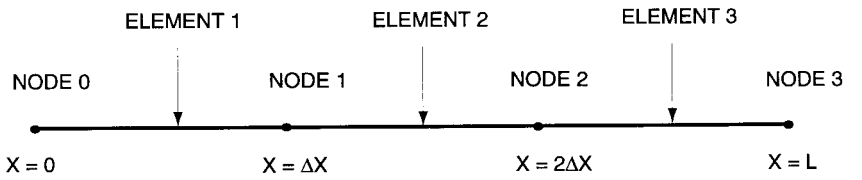

Figure 4. Domain for a simple test problem.

where $U \geq 0$ is the specified boundary value of u. A Neumann condition can be used at the right boundary to approximate a non-reflective boundary

$$\frac{\partial u(x = L, t)}{\partial x} = 0 \tag{17}$$

Global assembly of all the elements in the domain yields a 3×3 tridiagonal system of linear equations for the values of u at the global nodes, i.e.

$$\mathbf{Au} = \mathbf{b} \tag{18}$$

where $\mathbf{u}^T = (u_1, u_2, u_3)$, $\mathbf{b}^T = (-c_l U, 0, 0)$, and

$$\mathbf{A} = \begin{pmatrix} c_d & c_u & 0 \\ c_l & c_d & c_u \\ 0 & c_l & c'_d \end{pmatrix} \tag{19}$$

where

$$c_l = \frac{1}{6} + \frac{r}{2} - \frac{\alpha}{2}\left(r + \frac{1}{2}\right) \tag{20}$$

$$c_d = \frac{2}{3} + \alpha r \tag{21}$$

$$c_u = \frac{1}{6} - \frac{r}{2} - \frac{\alpha}{2}\left(r - \frac{1}{2}\right) \tag{22}$$

$$c'_d = \frac{1}{3} + \frac{r}{2} + \frac{\alpha}{2}\left(r + \frac{1}{2}\right) \tag{23}$$

This system can be solved for

$$u_1 = \frac{c_l c_u (c_d - c'_d) - c'_d (c_d^2 - c_l c_u)}{c'_d (c_d^2 - c_l c_u) - c_l c_d c_u} \frac{c_l U}{c_d} \tag{24}$$

$$u_2 = \frac{c_l^2 c'_d U}{c'_d (c_d^2 - c_l c_u) - c_l c_d c_u} \tag{25}$$

$$u_3 = \frac{-c_l^3 U}{c'_d (c_d^2 - c_l c_u) - c_l c_d c_u} \tag{26}$$

Equations (24)–(26) may be rewritten as

$$u_1 = C_1(\alpha, r) U \tag{27}$$

$$u_2 = C_2(\alpha, r) U \tag{28}$$

$$u_3 = C_3(\alpha, r) U \tag{29}$$

where the expressions for the C_i terms can be inferred from Equations (24)–(26). In order to maintain the monotonicity of the solution, the following conditions must be met:

$$0 \leq C_1,\ C_2,\ C_3 \leq 1 \tag{30}$$

$$C_1 \geq C_2 \geq C_3 \tag{31}$$

Figure 5 shows a series of plots of the C_i terms as a function of r for $\alpha = 1$. The spatial increment has been fixed such that $\Delta x = 1$. It can be seen that both conditions given by Equations (30) and (31) are met in this case. However, for the case when $\alpha = 0.75$, shown in Figure 6, it can be seen that for values of r larger than approximately 0.17, C_1 becomes negative. Therefore, the solution loses monotonicity and an undershoot develops at node 1. This is a serious problem, because in this particular case, the optimal value of r as suggested by Raymond and Garder is $1/\sqrt{15} \approx 0.258$. Therefore, the use of the optimal value of r will yield undershooting, but using $r = 0$ (Bubnov–Galerkin) will not. Figures 7 and 8 show the cases when $\alpha = 0.5$ and 0.25 respectively. The undershooting at node 1 becomes progressively worse as α is reduced. Note also that for these two cases, the undershooting at node 1 is present for all $r \geq 0$ and becomes progressively worse as r gets larger. Therefore, the undershoots will be larger for any non-zero r than for the Bubnov–Galerkin ($r = 0$) method. This is further illustrated by examining C_i as a function of α for the cases $r = 0$ and $r = 1/\sqrt{15}$ shown in Figures 9 and 10. It can be seen that the undershooting at node 1 is greater for the optimal choice of r than for the Bubnov–Galerkin solution. Undershooting also occurs over a larger range of α in the case when $r = 1/\sqrt{15}$.

The preceding example was presented for only one time step. In order to illustrate what happens as further time steps are taken, a numerical experiment has been performed where $u_L = 100$, $u_R = 0$, $a = 0.001$ m s^{-1}, $\Delta x = 0.1$ m, and $\Delta t = 1$ s. Thus, $\alpha = 0.01$, which should

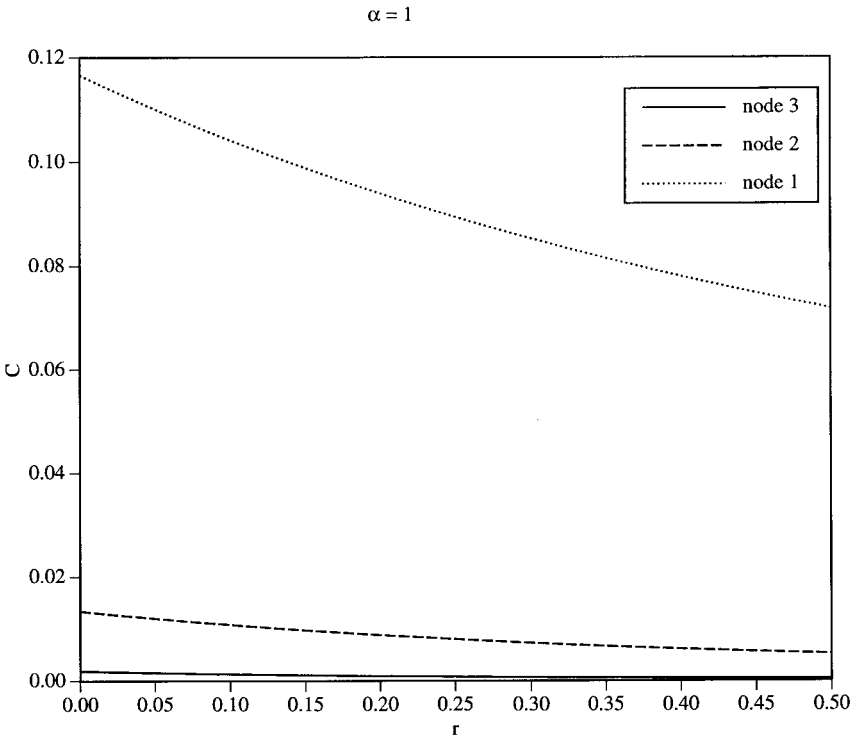

Figure 5. Plot of C_i as a function of r for $\alpha = 1$.

make $Q < 0$ at the base of the discontinuity. The Petrov–Galerkin solution is shown at two different times in Figure 11. At each time, undershooting is clearly visible, although the oscillations appear to be dampening with time. This is probably attributable to the accumulation of truncation errors and the subsequent smearing of the sharp wave. Equation (13) shows that Q is proportional to Δu, therefore, as Δu decays due to the numerical smearing of the wave, the magnitude of Q diminishes, creating a smaller sink.

The Bubnov–Galerkin solution to the same problem is also shown in Figure 11. Although there is still evidence of undershooting, it is much less pronounced than in the Petrov–Galerkin solution. Note that because the grid Peclet number is infinite for this problem, the Bubnov–Galerkin solution does develop oscillations at the top of the wave. These waves are damped in the Petrov–Galerkin solution as it was designed to do.

The same experiment as above is now performed except that now $\Delta t = 100$ s. Thus, $\alpha = 1$ and no undershooting should be expected. Figure 12 shows the Petrov–Galerkin solutions at the same two times. Clearly, undershooting is absent and in fact, the base of the wave is slightly smeared. The Bubnov–Galerkin solutions at the same two times are shown in Figure 12. Again there are no undershoots and the overshoots at the top of the wave are more severe than in the Petrov–Galerkin solutions.

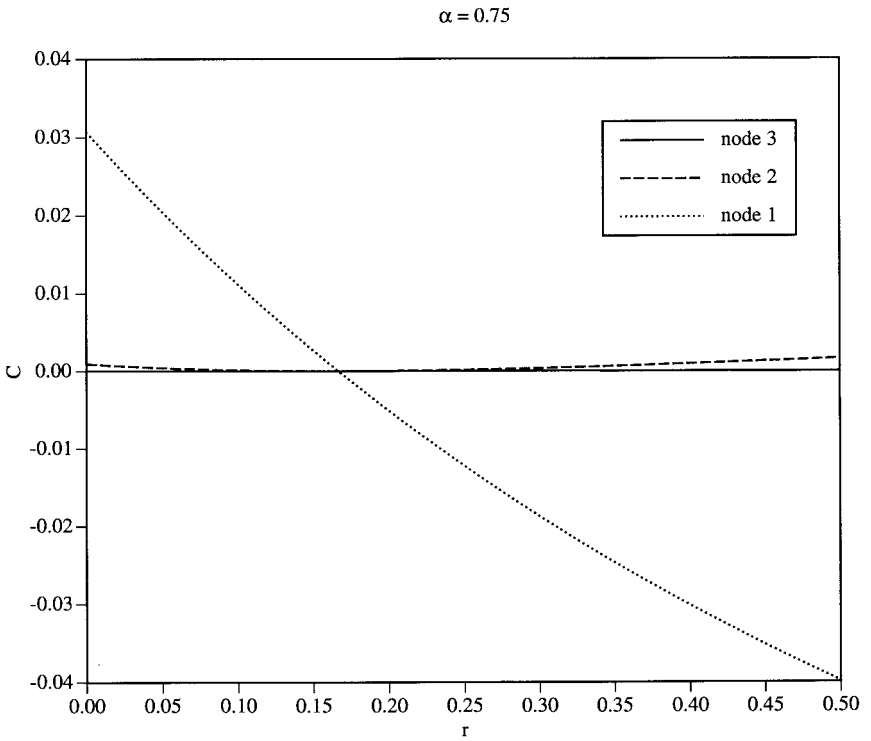

Figure 6. Plot of C_i as a function of r for $\alpha = 0.75$.

The observed smearing of the lower portion of the wave may be attributable to the increased truncation error present due to the use of a larger Δt. Therefore, the same experiment as shown in Figure 11 is performed except that now $a = 0.1$ m s^{-1}. Figure 13 shows the Petrov–Galerkin solutions for $\Delta t = 1$ s and 0.1 s. Undershooting is again present in the case where $\alpha < 1$. It is interesting to note that although the larger α helps to dissipate the undershoots, it appears to increase the overshoots at the top of the wave. The Bubnov–Galerkin solutions are presented in Figure 14. Again, the undershooting is less severe in this case and the overshoots are more severe when $\alpha = 1$.

3. FLOW IN UNSATURATED POROUS MEDIA

The governing equation to be examined in this case is the one-dimensional Richards' equation. It can be written as

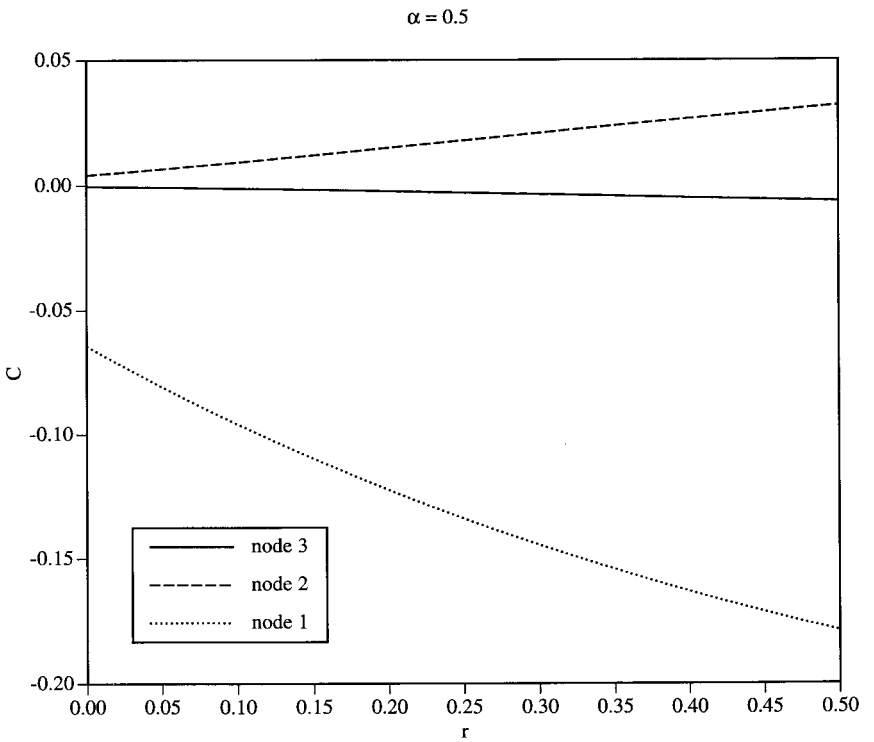

Figure 7. Plot of C_i as a function of r for $\alpha = 0.50$.

$$F\frac{\partial\psi}{\partial t} - \frac{\partial}{\partial x}\left(K\frac{\partial}{\partial x}(\psi + x)\right) = 0 \tag{32}$$

where ψ is the pressure head of the fluid in the porous medium, x denotes the vertical direction, and t represents time. Note that x is positive upwards and K is the hydraulic conductivity of the medium. The term F is defined as

$$F = S_s + \frac{d\theta}{d\psi} \tag{33}$$

where S_s is the specific storage coefficient, which includes the compressibility effects of both the porous matrix and fluid, and θ denotes the moisture content of the medium.

Substituting the following expression into Equation (32):

$$\frac{\partial K}{\partial x} = \frac{dK}{d\psi}\frac{\partial\psi}{\partial x} \tag{34}$$

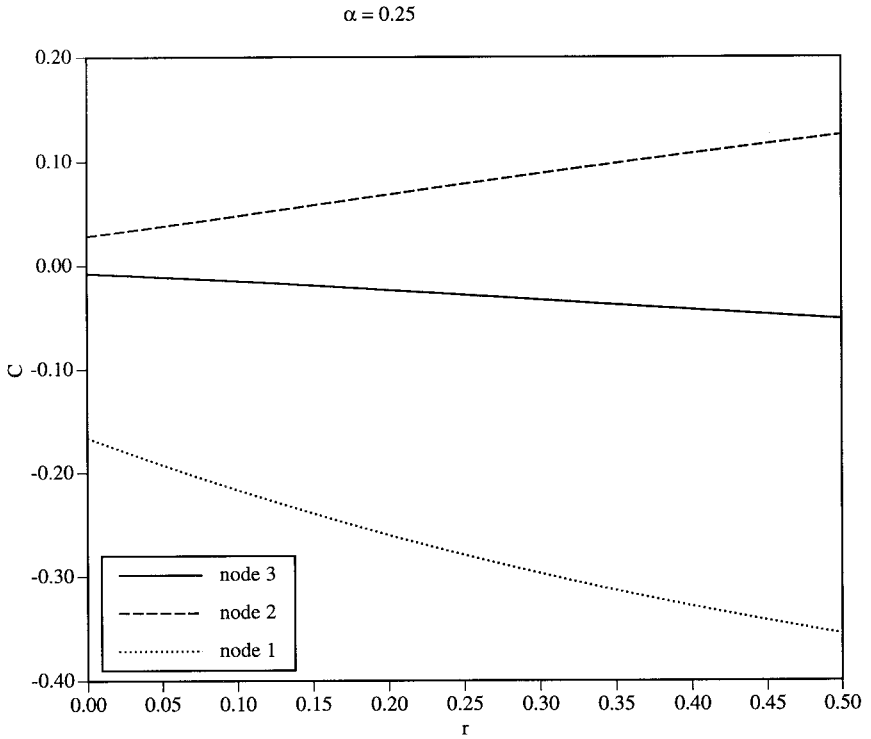

Figure 8. Plot of C_i as a function of r for $\alpha = 0.25$.

yields

$$F\frac{\partial\psi}{\partial t} - \frac{\mathrm{d}K}{\mathrm{d}\psi}\frac{\partial\psi}{\partial x} - \frac{\partial}{\partial x}\left(K\frac{\partial\psi}{\partial x}\right) = 0 \tag{35}$$

which resembles an advection–diffusion equation. In this form, it would seem that a Petrov–Galerkin approximation to the equation would be ideal. In fact, researchers have investigated solving this form of the equation [6] by utilizing some form of upwind differencing [7].

If the Petrov–Galerkin finite element discretization is applied to Equation (35), the resulting elemental equation is

$$M_{ij}\frac{\mathrm{d}\psi_j}{\mathrm{d}t} + S_{ij}\psi_j = 0 \tag{36}$$

in which

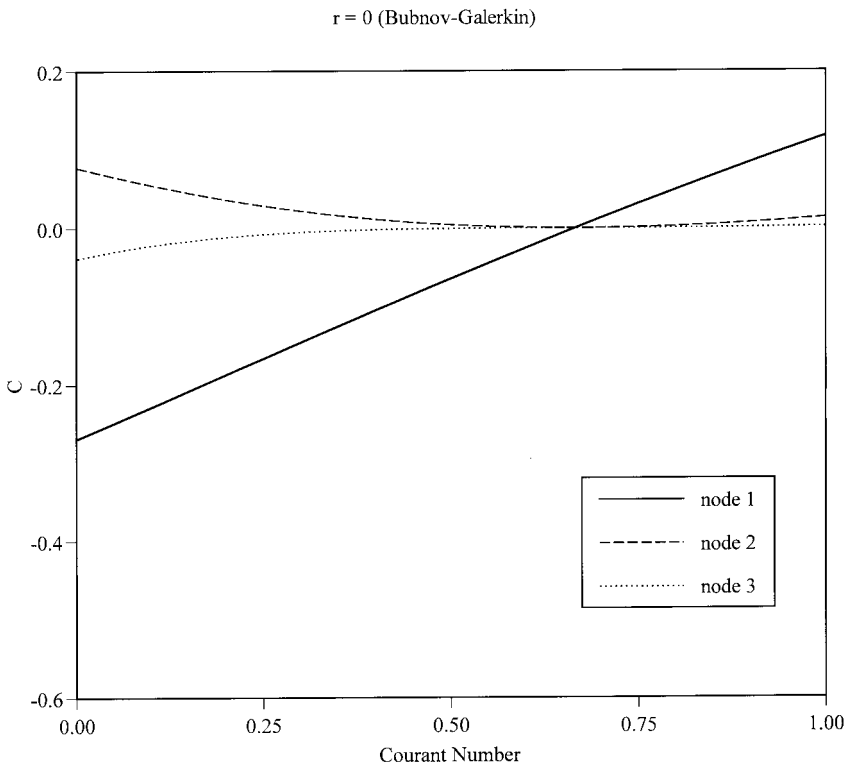

Figure 9. Plot of C_i as a function of α for $r = 0$.

$$M_{ij} = \int_{\Omega} F\left(N_i + p\frac{\partial N_i}{\partial x}\right) N_j \, \mathrm{d}\Omega \tag{37}$$

$$S_{ij} = \int_{\Omega} -\left(N_i + p\frac{\partial N_i}{\partial x}\right)\frac{\mathrm{d}K}{\mathrm{d}\psi}\frac{\partial N_j}{\partial x} + K\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}\,\mathrm{d}\Omega \tag{38}$$

where Ω denotes the domain of the element. Time marching is again accomplished through the standard Crank–Nicolson method, i.e.

$$\frac{\mathrm{d}\psi_j}{\mathrm{d}t} \approx \frac{\psi_j^{n+1} - \psi_j^n}{\Delta t} \tag{39}$$

$$\psi_j \approx \frac{1}{2}(\psi_j^{n+1} + \psi_j^n) \tag{40}$$

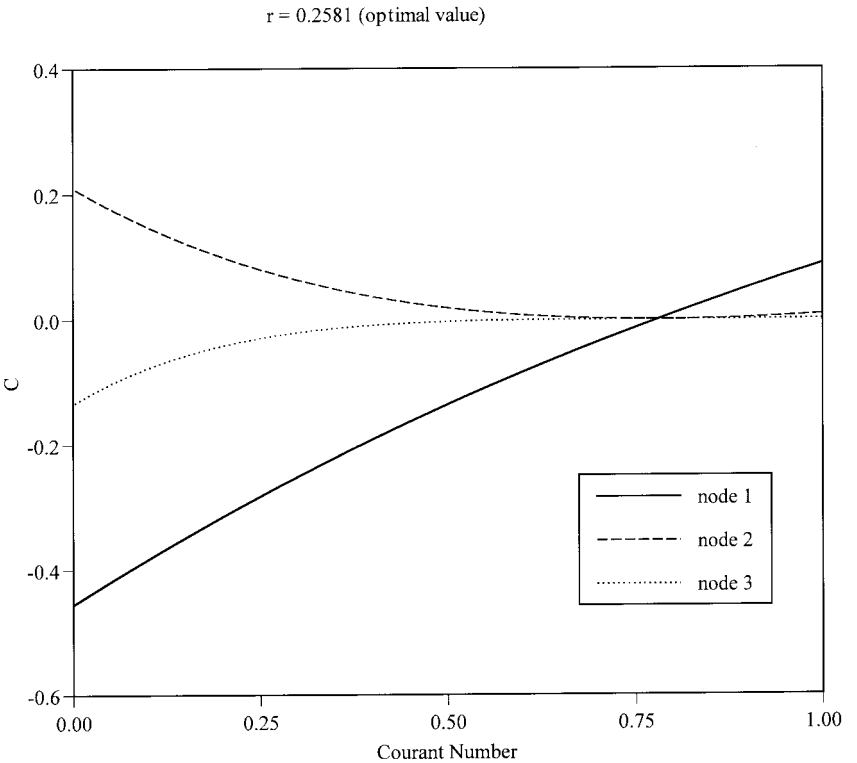

Figure 10. Plot of C_i as a function of a for $r = 1/\sqrt{15}$ (optimal values).

The preceding expression for Q_j, given by Equation (13), can be extended to include the natural diffusion of this problem, i.e.

$$Q_j \approx \frac{|r\alpha \Delta u|}{2\Delta t}\left(|\alpha| - 1 + \frac{\beta}{|r|}\right) \tag{41}$$

where the Courant number is now defined as

$$\alpha = -\frac{\mathrm{d}K}{\mathrm{d}\psi}\frac{\Delta t}{F\Delta x} \tag{42}$$

and $\beta = K\Delta t/(F\Delta x^2)$. Thus, the inclusion of natural diffusion has somewhat relaxed the condition to prevent undershooting by allowing α to be smaller before Q_j becomes negative.

Some numerical experiments have been performed involving a domain of a homogeneous, sandy soil, which is 40 cm long and a specified $\psi = -10$ cm at the top boundary and $\partial\psi/\partial x = 0$ at the bottom boundary. Initially, $\psi = -100$ cm. The hydraulic conductivity is

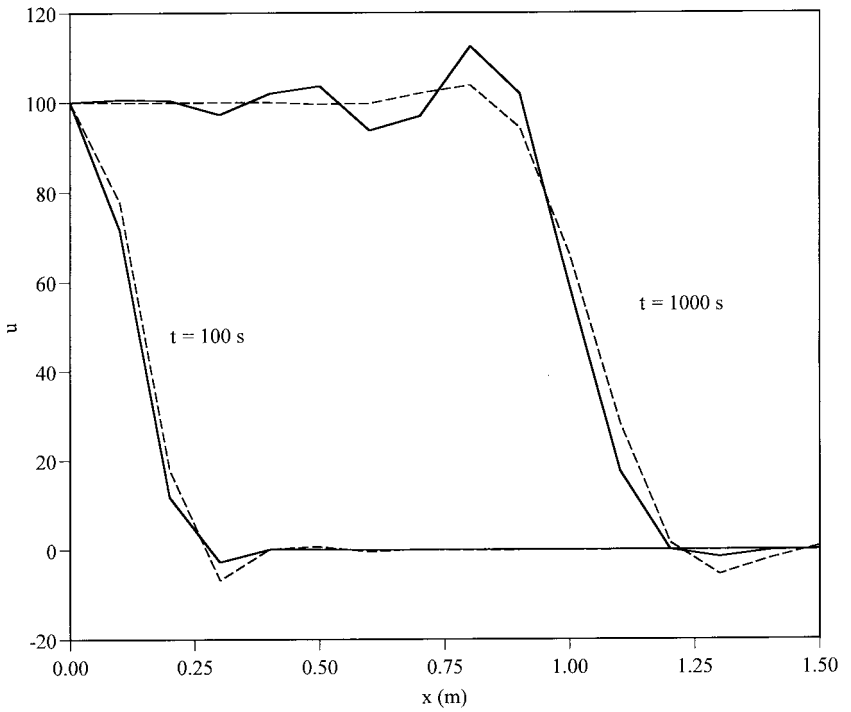

Figure 11. Bubnov and Petrov–Galerkin solutions at $t = 100$ and 1000 s for scalar advection with $\alpha = 0.01$. Solid and dashed lines correspond to the Bubnov and Petrov–Galerkin methods respectively.

defined as $K = K_{\text{sat}} k_{\text{rw}}$, where K_{sat} is the saturated hydraulic conductivity and k_{rw} is the relative permeability. A value $K_{\text{sat}} = 34$ cm h^{-1} is used in all experiments. The remaining soil properties are given as

$$k_{\text{rw}} = \frac{c_1}{c_1 + |\psi|^{c_2}} \tag{43}$$

$$\theta = \frac{c_3(\theta_{\text{s}} - \theta_{\text{r}})}{c_3 + |\psi|^{c_4}} + \theta_{\text{r}} \tag{44}$$

where $c_1 = 1.175 \times 10^6$, $c_2 = 4.74$, $c_3 = 1.611 \times 10^6$, and $c_4 = 3.96$, are empirical constants. The terms θ_{s} and θ_{r} represent the saturated and residual water contents of the sand and are equal to 0.95 and 0.25 respectively. In addition, S_{s} is assumed to be negligible.

Figure 15 shows a comparison of the Bubnov and Petrov–Galerkin solutions at $t = 5$ and 100 s using $\Delta x = 1$ cm and $\Delta t = 0.5$ s. Undershooting is clearly visible in the solutions despite the presence of natural diffusion, and again is more pronounced in the Petrov–Galerkin

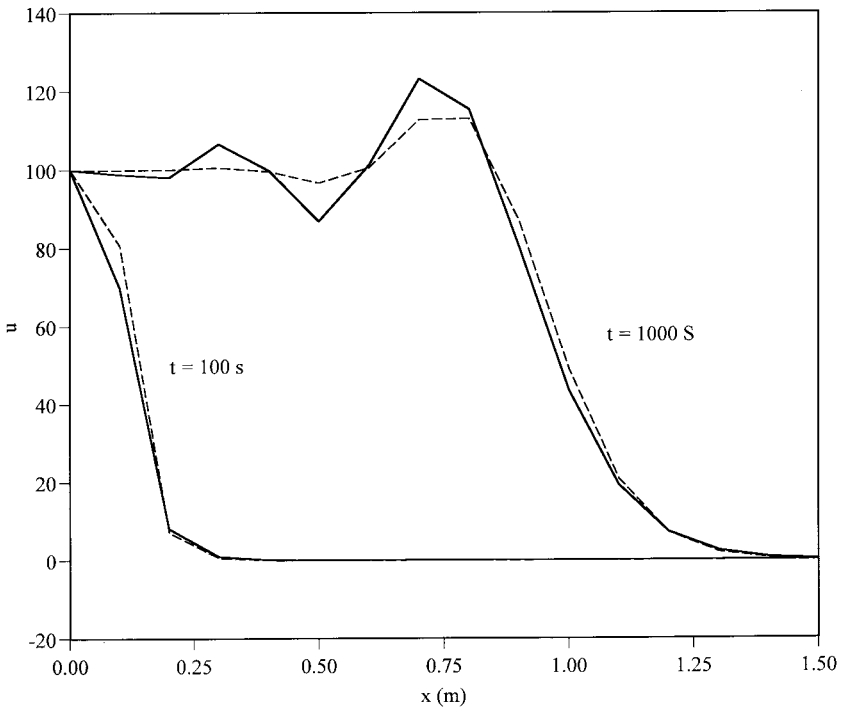

Figure 12. Bubnov and Petrov–Galerkin solutions at $t = 100$ and 1000 s for scalar advection with $\alpha = 1$. Solid and dashed lines correspond to the Bubnov and Petrov–Galerkin methods respectively.

solutions. The undershoots appear to dissipate as the wave front is smeared as in the linear advection case. The ineffectiveness of the natural diffusion to alleviate the oscillations is attributable to the unique non-linearity of K in this problem. This simulation involves the seepage of water into a dry soil, and therefore in front of the wave K is very small and thus provides little help in smoothing the oscillations. This can be seen in Figure 16, which shows a plot of α and β as a function of x at $t = 100$ s. Near the wetting front, it can be seen that β and α are both quite small.

Equation (41) shows that decreasing Δx or increasing Δt will yield $Q_j \geq 0$. Increasing Δt, however, increases the truncation error and is therefore not a desirable option. For this particular problem, $\Delta x \leq 0.22$ cm would be required in order to make $Q_j \geq 0$. The choice of $\Delta x = 0.2$ cm does in fact eliminate the undershoots, however, it also introduces oscillations at the specified pressure head boundary. These can be eliminated with either a reduction in Δt or the use of fully implicit time stepping. The former option reintroduces the undershoots because it reduces α. The latter choice introduces larger truncation errors, which may lead to mass balance errors.

Despite these problems, it is seen that there are no oscillations near the top of the wave for either the Petrov or the Bubnov–Galerkin solutions, which indicates that high frequency

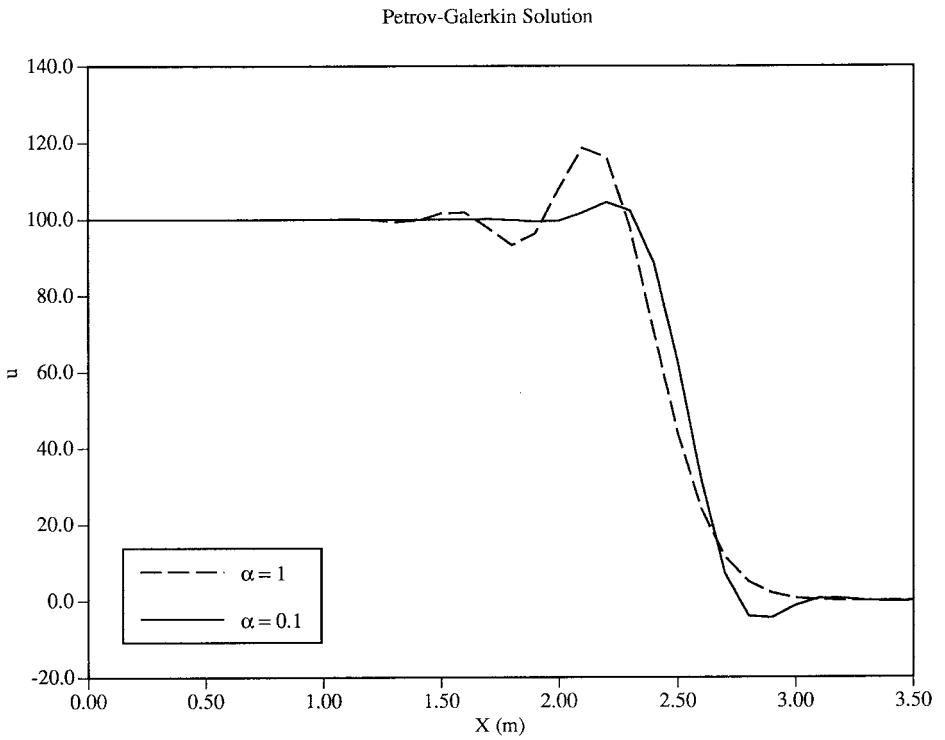

Figure 13. Petrov–Galerkin solutions for scalar advection at $t = 25$ s with $\alpha = 1$ and 0.1.

oscillatory waves that are well damped by the Petrov–Galerkin method are absent in this simulation. In addition, the undershoots are aggravated by Petrov–Galerkin upwinding. In fact, this problem is diffusion-dominated in the sense that the grid Peclet number, $Pe = a\Delta x/K$, is less than 1. The Peclet number is plotted in Figure 17 as a function of x for $\Delta x = 1$ and 0.2 cm Thus, Petrov–Galerkin upwinding is probably not warranted for this problem.

However, it would still be desirable to dampen the oscillations since they can cause problems. For example, if the transport of sediment or a contaminant is modeled with the advection–diffusion equation, undershoots may yield negative concentrations. This is not only physically incorrect but may lead to numerical instabilities if the computed concentrations are needed as input into other models. The previously presented Petrov–Galerkin method works well in damping the overshoots at discontinuities, but fails to dampen the undershoots, and actually makes them worse in some instances. Therefore, an alternative method of model design is considered in which the monotonicity of the solution is preserved.

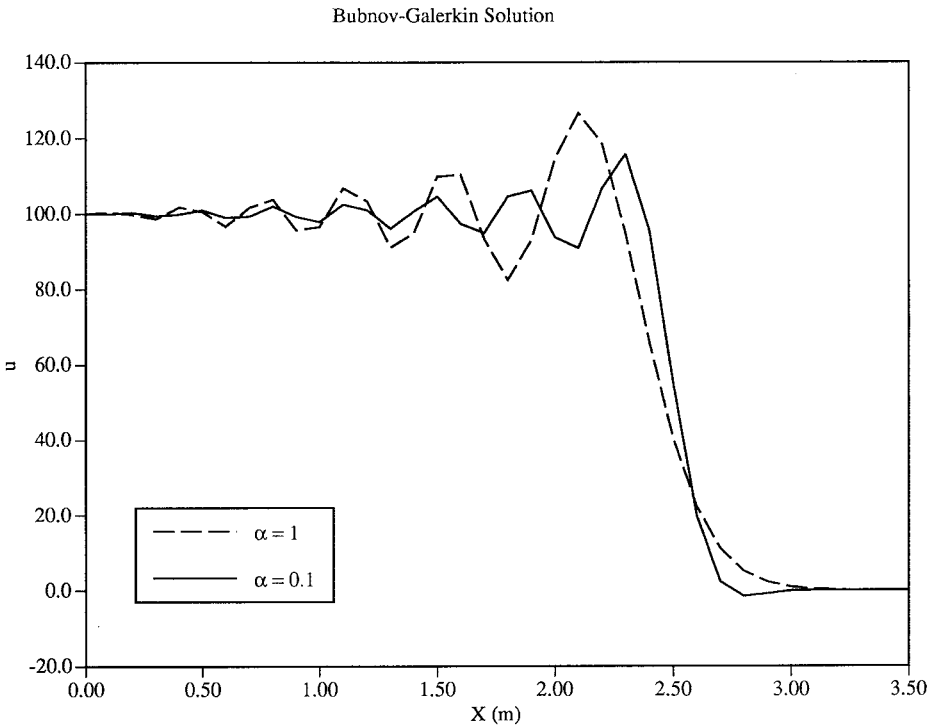

Figure 14. Bubnov–Galerkin solutions for scalar advection at $t = 25$ s with $\alpha = 1$ and 0.1.

4. MONOTONICITY PRESERVATION

Godunov's theorem states that any linear scheme that is greater than first-order accurate will not preserve the monotonicity of the solution [8]. The Petrov and Bubnov–Galerkin methods are examples of such techniques. As previously mentioned, the Petrov–Galerkin method is optimally designed to minimize the dispersive error while simultaneously introducing selective dissipation to dampen high frequency waves. Alternatively, the scheme could be designed such that the monotonicity of the solution is preserved. Such an analysis has in fact been performed by Bøe [5] for the quasi-linear advection–diffusion equation. However, while he derived monotonicity conditions from a finite element perspective, here the high resolution, TVD finite volume method is examined in order to illustrate the link between the Petrov–Galerkin and TVD finite volume method.

The TVD finite volume method can be summarized as follows. First, compute an average slope of u in cell j, i.e.

$$\bar{s}_j = \text{avg}(s_{j+1}, s_{j+1/2}, s_{j-1/2}) \tag{45}$$

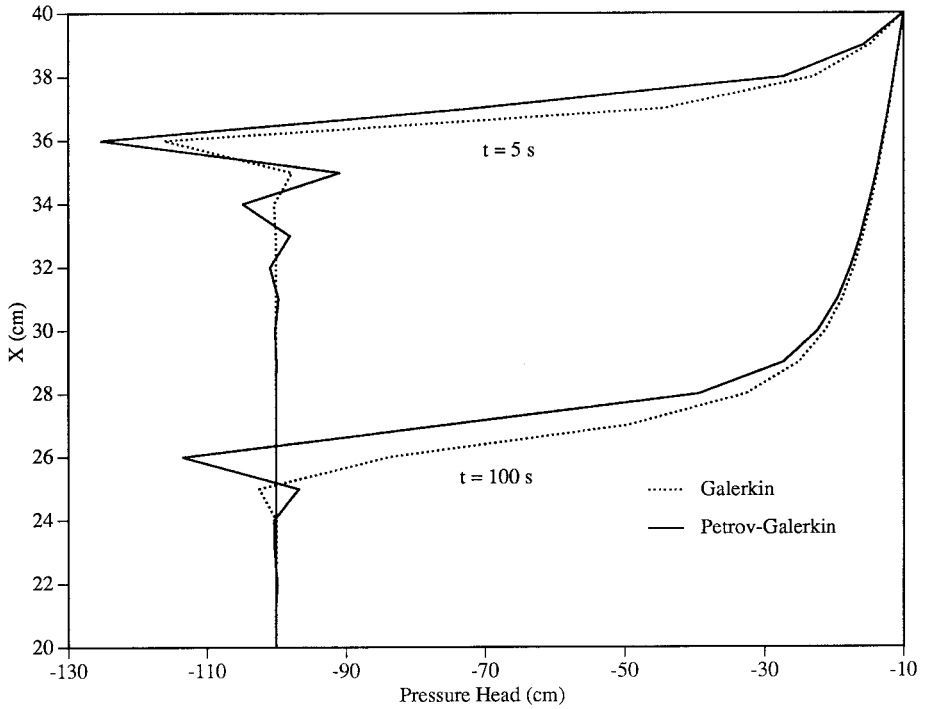

Figure 15. Comparison of Petrov and Bubnov–Galerkin solutions with Richards' equation at $t = 5$ and 100 s.

where $s_{j+1/2} = u_{j+1} - u_j$ and $s_{j-1/2} = u_j - u_{j-1}$. The average is chosen such that the monotonicity of the solution is preserved. For example, the Superbee average [9] is

$$\bar{s}_j = \begin{cases} 0, & s_{j+1/2}s_{j-1/2} \le 0 \\ \min(\max(s_{j+1/2}, s_{j-1/2}),\ \min(2s_{j+1/2}, 2s_{j-1/2})), & s_{j+1/2}s_{j-1/2} > 0 \end{cases} \tag{46}$$

There exist many other choices available in the literature, each of which has different dissipative properties. The Superbee limiter happens to be one of the least dissipative limiters available, which makes it ideal for capturing sharp discontinuities. The next step is to compute the predictor solution in cell j at $\Delta t/2$

$$u_j^{n+1/2} = u_j^n - \frac{\alpha}{2}\bar{s}_j \tag{47}$$

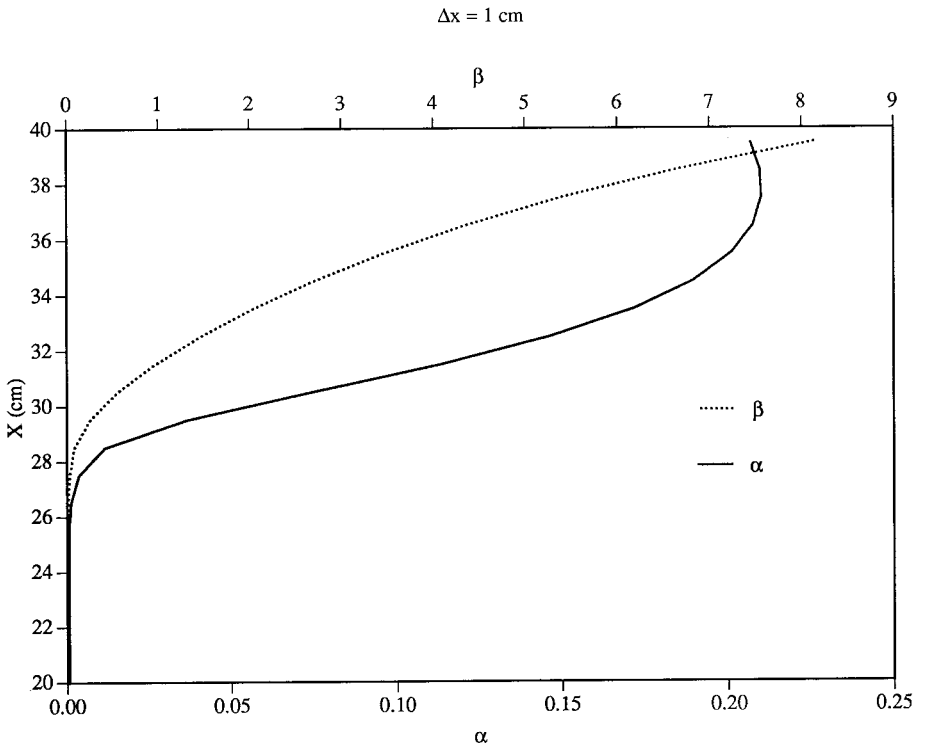

Figure 16. Plot of α and β at $t = 100$ s for the problem shown on Figure 15.

which is done to achieve second-order time accuracy. Then perform a linear reconstruction to the cell faces, denoted $j \pm 1/2$

$$u^{\mathrm{L}}_{j+1/2} = u^{n+1/2}_{j} + \frac{1}{2}\bar{s}_j \tag{48}$$

$$u^{\mathrm{R}}_{j+1/2} = u^{n+1/2}_{j+1} - \frac{1}{2}\bar{s}_{j+1} \tag{49}$$

$$u^{\mathrm{L}}_{j-1/2} = u^{n+1/2}_{j+1} + \frac{1}{2}\bar{s}_{j-1} \tag{50}$$

$$u^{\mathrm{R}}_{j-1/2} = u^{n+1/2}_{j} - \frac{1}{2}\bar{s}_j \tag{51}$$

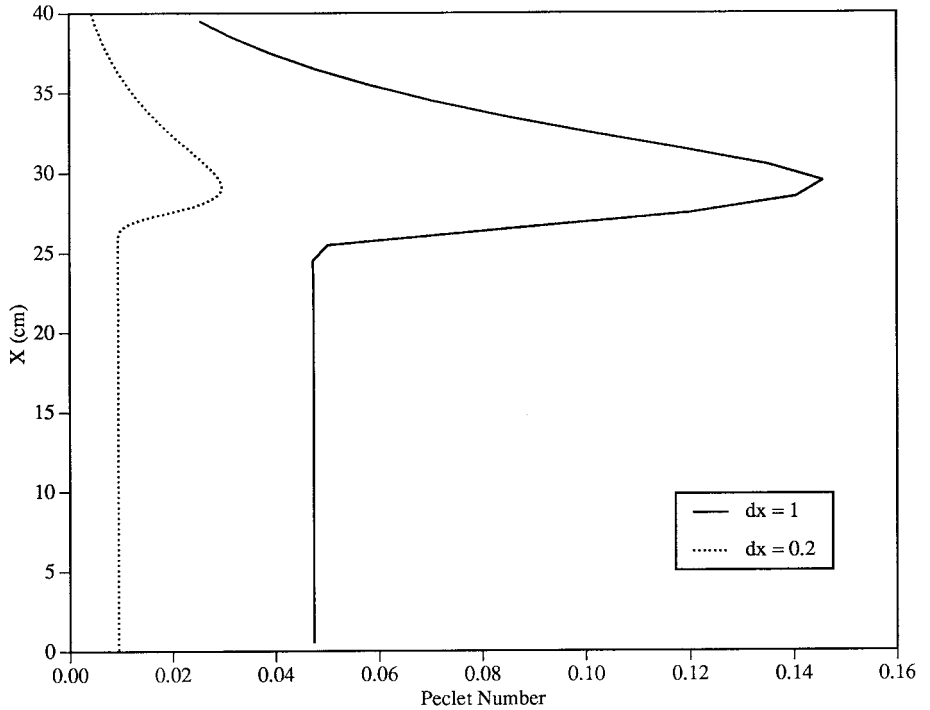

Figure 17. Plot of the Peclet number at $t = 100$ s for the problem shown in Figure 15 with $\Delta x = 1$ and 0.2 cm.

which is also called a monotone upstream scheme for conservation laws (MUSCL) reconstruction and was developed by van Leer [10]. The superscripts L and R denote the values to the left and right of the face respectively. This is where a limited amount of anti-dissipation is added to achieve second-order spatial accuracy. The same feature can also be found in the well-known Lax–Wendroff method. However, because this method utilizes a linear average of the data, i.e., $s_j = u_{j+1} - u_j$, it does not limit the amount of anti-dissipation added, which gives rises to oscillatory solutions.

With the computed values at the interfaces known, the fluxes at the cell interfaces are computed using Roe's approximate Riemann solver [11]. For the one-dimensional, scalar advection equation, this technique is equivalent to the upwind method, which can be written as

$$f_{j+1/2}^{n+1/2} = \frac{1}{2}\left(f(u_{j+1/2}^{\mathrm{L}}) + f(u_{j+1/2}^{\mathrm{R}})\right) - \frac{|\alpha|\Delta x}{2\Delta t}\left(u_{j+1/2}^{\mathrm{R}} - u_{j+1/2}^{\mathrm{L}}\right) \tag{52}$$

$$f_{j-1/2}^{n+1/2} = \frac{1}{2}\left(f(u_{j-1/2}^{\mathrm{L}}) + f(u_{j-1/2}^{\mathrm{R}})\right) - \frac{|\alpha|\Delta x}{2\Delta t}\left(u_{j-1/2}^{\mathrm{R}} - u_{j-1/2}^{\mathrm{L}}\right) \tag{53}$$

where $f(u) = au$ is the advective flux. The final step is to evolve the data

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{f_{j+1/2}^{n+1/2} - f_{j-1/2}^{n+1/2}}{\Delta x} = 0 \tag{54}$$

The complete scheme can be summarized as

$$\begin{aligned} u_j^{n+1} - u_j^n = & -\frac{\alpha}{2}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{|\alpha|}{2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + \frac{\alpha(\alpha-1)}{4}\left(\bar{s}_{j+1} - \bar{s}_{j-1}\right) \\ & - \frac{|\alpha|(\alpha-1)}{4}\left(\bar{s}_{j+1} - 2\bar{s}_j + \bar{s}_{j-1}\right) \end{aligned} \tag{55}$$

The Petrov–Galerkin scheme with a lumped mass matrix and fully explicit time stepping can be written as

$$u_j^{n+1} - u_j^n = -\frac{\alpha}{2}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{\alpha p}{\Delta x}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) \tag{56}$$

Bøe [5] demonstrated that mass lumping is needed to preserve monotonicity. When $p = |\alpha|\Delta x/(2\alpha)$, the standard upwind difference scheme is recovered, while selecting $p = a\Delta t/2$ yields the Lax–Wendroff scheme. In order to recover the TVD finite volume method, p should be computed as

$$p = \frac{\Delta x}{2}\left\{\mathrm{sign}(\alpha) + \frac{\alpha - 1}{2\Delta^2 u}\left(\Delta s - \mathrm{sign}(\alpha)\Delta^2 s\right)\right\} \tag{57}$$

where

$$\Delta^2 u = u_{j+1}^n - 2u_j^n + u_{j-1}^n \tag{58}$$

$$\Delta s = \bar{s}_{j+1} - \bar{s}_{j-1} \tag{59}$$

$$\Delta^2 s = \bar{s}_{j+1} - 2\bar{s}_j + \bar{s}_{j-1} \tag{60}$$

Thus, the level of dissipation is not merely grid-dependent, but also varies with the solution. This makes the scheme non-linear and therefore Godunov's theorem is not violated. Note, however, the spatial stencil has been increased from three to five nodes by including nodes $j-2$ and $j+2$. This is needed to properly monitor the smoothness of the solution in order to discriminate between numerical and physical extrema. A three-node stencil would not allow for such discrimination.

Figure 18 shows a comparison of the Petrov–Galerkin solutions with $p = \Delta x/\sqrt{15}$ and no mass lumping (labeled consistent Petrov–Galerkin) and p from Equation (57) with mass lumping (labeled Superbee) for the same problem as shown in Figures 13 and 14. The Superbee limiter was utilized for constructing the scalar gradients. From the plot, it can be seen that for the monotone Petrov–Galerkin method (labeled Superbee), all oscillations have been damped while maintaining the sharpness of the discontinuity.

One may be tempted to simply lump the mass matrix in an existing finite element model in an attempt to eliminate the oscillations. Note that for the Petrov–Galerkin method, doing so eliminates the mixed time–space derivative term in Equation (6), and thus eliminates the undershooting problem. However, this leaves only the artificial dissipation term, which leads to an overly smeared solution, that is in fact only mildly less dissipative than the first-order upwind scheme, as shown in Figure 18. This can also be demonstrated by a Fourier analysis. Figure 18 also shows the lumped Bubnov–Galerkin solution. Although the undershoots have been eliminated, the overshoots at the top of the wave are still present. Therefore, mass lumping alone is not sufficient to preserve the monotonicity of the solution.

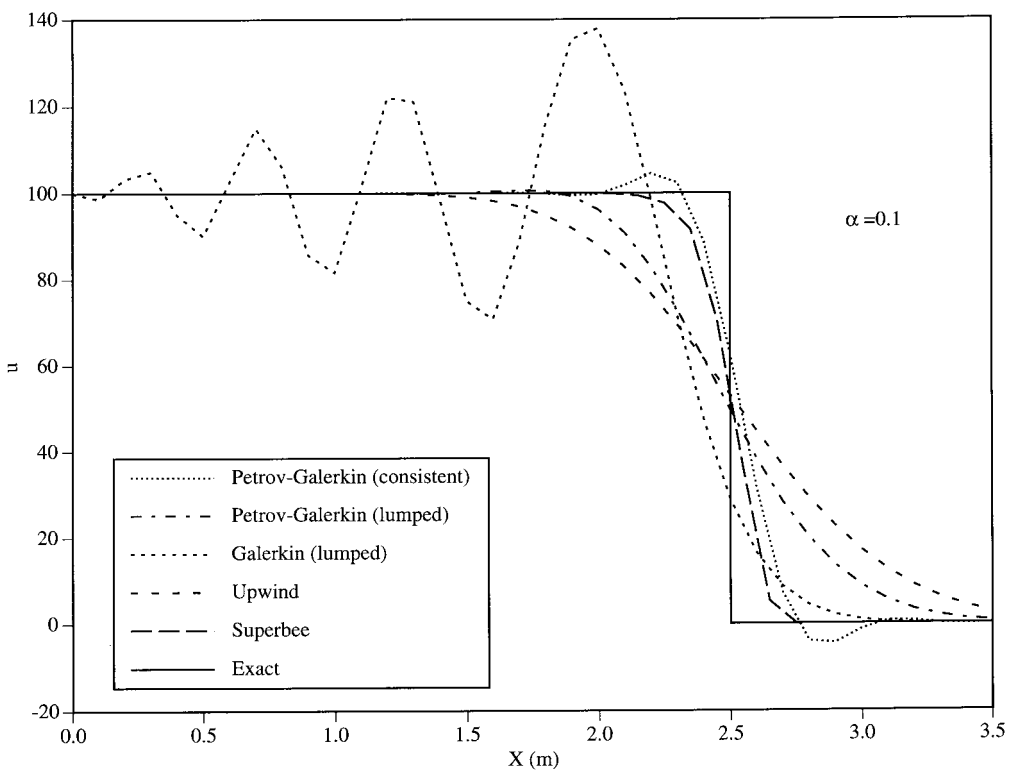

Figure 18. Comparison of various methods with the exact solution to the scalar advection equation for the problem shown in Figures 13 and 14.

5. CONCLUSIONS

It has been demonstrated that for $\alpha < 1$, the Petrov–Galerkin method is actually less dissipative than the standard Bubnov–Galerkin scheme when considering scalar advection. When the grid Peclet number is larger or infinite, the Petrov–Galerkin scheme damps oscillations in the Bubnov–Galerkin solution at the top of discontinuities quite effectively. However, it also introduces undershoots that are more severe than in the Bubnov–Galerkin solution. This is contrary to a standard Fourier analysis, which shows the Petrov–Galerkin scheme is always more dissipative than the Bubnov–Galerkin method. However, the undershoots do not appear to grow with time and therefore seem stable.

The analysis has been extended to the advection–diffusion equation and it was shown that the presence of diffusion does indeed help to minimize the undershoots. However, it was also demonstrated that the non-linear behavior of K in Richards' equation prevents the presence of natural diffusion from alleviating the undershooting phenomenon. Also, the small grid Peclet number and the lack of overshooting in the Bubnov–Galerkin solution indicate an absence of high frequency waves that the Petrov–Galerkin method is designed to dissipate. Thus, Petrov–Galerkin upwinding is probably not warranted in this problem, but should only be used in advection-dominated flows and not just because solution undershoots are present. When upwinding is warranted, the condition that $\alpha \approx 1$ should be enforced everywhere in the domain will help to alleviate the undershooting problem and maintain the overall accuracy of the solution for time-dependent problems. This introduces an unfortunate tradeoff since as α increases, the overshoots at the top of discontinuities in both the Bubnov and Petrov–Galerkin solutions increase in magnitude.

Alternatively, a method could be designed that specifically preserves the monotonicity of the solution. Finite element methods designed with criteria based upon the minimization of the phase or amplitude errors yield linear schemes that violate Godunov's theorem. It was demonstrated that a Petrov–Galerkin method can be constructed such that it is equivalent to the high resolution, TVD finite volume method widely used in the field of gas dynamics. Thus, with the simple addition of a sub-routine to compute non-linear averages of the data, a monotone Petrov–Galerkin method can be constructed that captures discontinuities sharply without the numerical oscillations of the standard Pertrov–Galerkin method. However, such a monotone method requires the lumping of the mass matrix and the use of fully explicit time stepping. Therefore, the method is only stable for $\alpha \leq 1$. In addition, maintaining the monotonicity of the solution requires the enlargement of the spatial stencil and introduces additional complications regarding the specification of boundary conditions.

It should be mentioned that the monotone Petrov–Galerkin eliminates all oscillations at discontinuities, while the standard Petrov–Galerkin method merely dampens them. Therefore, it must be decided if the oscillations present in the standard Petrov–Galerkin solution are serious enough to warrant the construction of a monotone scheme. In particular, the undershoots can have a disastrous impact on a simulation if u is supposed to be a purely positive quantity, such as the concentration of a sediment or chemical. Finally, the extension of the monotone Petrov–Galerkin scheme to two and three spatial dimensions is not straightforward. In two dimensions, for example, the finite element stencil contains nodes that are not used in the standard finite volume or finite difference stencils. This can be overcome

by choosing weighting functions other than the bilinear functions frequently used. More research is needed to further quantify this.

APPENDIX A. NOMENCLATURE

a advection velocity
f flux
j node index
K hydraulic conductivity
M_{ij} elements of the mass matrix
n time level
N_i linear trial functions
p Petrov–Galerkin parameter
Q source term
r Pertov–Galerkin parameter ($=p/\Delta x$)
S_{ij} elements of the stiffness matrix
t time
u scalar quantity
W_i weighting function
x spatial co-ordinate

Greek letters
α Courant number
ψ pressure head
θ moisture content
ω wave frequency
Δt time step
Δx nodal spacing
Ω finite element domain

REFERENCES

1. Hughes TJR, Brooks A. A theoretical framework for Petrov–Galerkin methods with discontinuous weighting functions: application to the streamline–upwind procedure. *Finite Elements in Fluids* 1982; **4**: 47–65.
2. Katopodes ND. A dissipative Galerkin scheme for open-channel flow. *Journal of Hydraulic Engineering* 1984a; **110**: 450–466.
3. Katopodes ND. Fourier analysis of dissipative FEM channel flow model. *Journal of Hydraulic Engineering* 1984b; **110**: 927–944.
4. Raymond WH, Garder A. Selective damping in a Galerkin method for solving wave problems with variable grids. *Monthly Weather Review* 1976; **104**: 1583–1590.
5. Bøe Ø. A monotone Petrov–Galerkin method for quasilinear parabolic differential equations. *SIAM Journal on Scientific Computing* 1993; **14**: 1057–1071.

6. El Kadi AI, Ling G. The Courant and Peclet number criteria for the numerical solution of the Richard's equation. *Water Resources Research* 1993; **29**: 3485–3494.
7. Huang K, Zhang R, van Genuchten MT. A Eularian–Lagrangian approach with an adaptively corrected method of characteristics to simulate variably saturated water flow. *Water Resources Research* 1994; **30**: 499–507.
8. Hirsch C. *Numerical Computation of Internal and External Flows*. Wiley: New York, 1990.
9. Sweby PK. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM Journal of Numerical Analysis* 1984; **21**: 995–1011.
10. van Leer B. Towards the ultimate conservative difference scheme. V. A second order sequel to Godunov's method. *Journal of Computational Physics* 1979; **32**: 101–136.
11. Roe PL. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics* 1981; **43**: 357–372.

