WATER RESOURCES RESEARCH, VOL. 26, NO. 7, PAGES :483–1496, JULY 1990

A General Mass-Conservative Numerical Solution for the Unsaturated Flow Equation

MICHAEL A. CELIA AND EFTHIMIOS T. BOULOUTAS

Water Resources Program, Department of Civil Engineering and Operations Research, Princeton University, Princeton, New Jersey

REBECCA L. ZARBA

Camp Dresser and McKee Incorporated, Boston, Massachusetts

Numerical approximations based on different forms of the governing partial differential equation can lead to significantly different results for unsaturated flow problems. Numerical solution based on the standard h-based form of Richards equation generally yields poor results, characterized by large mass balance errors and erroneous estimates of infiltration depth. Conversely, numerical solutions based on the mixed form of Richards equation can be shown to possess the conservative property, so that mass is perfectly conserved. This leads to significant improvement in numerical solution performance, while requiring no additional computational effort. However, use of the mass-conservative method does not guarantee good solutions. Accurate solution of the unsaturated flow equation also requires use of a diagonal time (or mass) matrix. Only when diagonal time matrices are used can the solution be shown to obey a maximum principle, which guarantees smooth, nonoscillatory infiltration profiles. This highlights the fact that proper treatment of the time derivative is critical in the numerical solution of unsaturated flow.

INTRODUCTION

Prediction of fluid movement in unsaturated soils is an important problem in many branches of science and engineering. These include soil science, agricultural engineering, environmental engineering, and groundwater hydrology. In virtually all studies of the unsaturated zone, the fluid motion is assumed to obey the classical Richards equation [*Hillel*, 1980]. This equation may be written in several forms, with either pressure head $h[L]$ or moisture content θ $[L^3/L^3]$ as the dependent variable. The constitutive relationship between θ and h allows for conversion of one form of the equation to another.

Three standard forms of the unsaturated flow equation may be identified: the "h-based" form, the "θ-based" form, and the "mixed form." These equations are written as

h based

$$C(h)\frac{\partial h}{\partial t} - \nabla\cdot K(h)\nabla h - \frac{\partial K}{\partial z} = 0 \qquad (1)$$

θ based

$$\frac{\partial \theta}{\partial t} - \nabla\cdot D(\theta)\nabla\theta - \frac{\partial K}{\partial z} = 0 \qquad (2)$$

Mixed

$$\frac{\partial \theta}{\partial t} - \nabla\cdot K(h)\nabla h - \frac{\partial K}{\partial z} = 0 \qquad (3)$$

In (1)–(3), $C(h) \equiv d\theta/dh$ is the specific moisture capacity function $[1/L]$, $K(h)$ is the unsaturated hydraulic conductivity $[L/T]$, $D(\theta) \equiv K(\theta)/C(\theta)$ is the unsaturated diffusivity $[L^2/T]$, z denotes the vertical dimension, assumed positive upward, and the porous medium is assumed to be isotropic. It is also assumed that appropriate constitutive relationships between θ and h and between K and h (or K and θ) are available.

Because these equations are nonlinear, analytical solution is not possible except for special cases. Therefore numerical approximations are typically used to solve the unsaturated flow equation. The standard approximations that are applied to the spatial domain are the finite difference method and the finite element method. These are usually coupled with a simple one-step Euler time-marching algorithm. For any Euler method other than the fully explicit forward method, nonlinear algebraic equations result and some linearization and/or iteration procedure must be used to solve the discrete equations. Standard iteration techniques include Picard and Newton methods.

This paper investigates the numerical behavior of standard approximation methods for the unsaturated flow equation. Solution using the h-based formulation and a backward Euler time discretization is shown to produce unacceptably large mass balance errors for many example calculations. This is true for any iteration method (Picard, Newton-Raphson, etc.). It is also true for both finite difference and finite element approximations in space, although finite elements are generally inferior to finite differences. Because use of the h-based formulation with a simple time-stepping method is widespread, these findings appear to have significant practical implications.

A modified numerical approach is proposed that alleviates the mass balance problems discussed above. This approach is based on a fully implicit (backward Euler) time approximation applied to the mixed form of the unsaturated flow equation. Proper expansion of the time derivative produces a simple computational algorithm that is perfectly mass conservative for numerical approximations that preserve

Paper number 90WR00196.
0043-1397/90/90WR-00196$05.00

spatial symmetry. Thus the finite difference and (Galerkin) finite element approximations using this mixed formulation are perfectly mass conservative. This approach is shown to be superior to the standard *h*-based approximations while requiring no more computational effort. However, conservation of mass is shown to be insufficient to guarantee good numerical solutions. For infiltration into dry soils, finite element approximations produce oscillatory solutions even while conserving mass. It is shown that diagonalization of the time matrix, which occurs naturally in finite difference approximations, is necessary and sufficient to guarantee nonoscillatory solutions. This demonstrates the importance of mass lumping in finite element solutions to unsaturated flow problems.

The paper begins with an overview of previous numerical approximations for unsaturated flow. This is followed by a brief explanation of the numerical methods used herein. The mass balance problems exhibited by *h*-based solutions are demonstrated via several example numerical calculations. The mixed equation approximation is then presented and shown to possess the conservative property for both finite difference and finite element spatial approximations. Numerical results are presented to illustrate the differences between the conservative mixed schemes and the *h*-based methods. Next the question of mass lumping in finite element approximations is addressed. In particular, diagonalization of the time matrix is shown to be necessary for monotonic solutions. This is followed by a general discussion of unsaturated flow modeling, including some effective ways to improve *h*-based solutions, and the extension of our results to problems in multiple dimension.

BACKGROUND

Numerical simulation of unsaturated flow has a significant history in the fields of soil science and groundwater hydrology. General overviews and thorough reviews of the literature may be found in the recent works of *Nielsen et al.* [1986] and *Milly* [1988]. Almost all unsaturated flow simulations use either the *h*-based form of Richards equation or the *θ*-based form. A variety of finite difference and finite element solution techniques have been used with each of these equation forms [*Neuman*, 1973; *Narasimhan and Witherspoon*, 1976; *Haverkamp et al.*, 1977; *Hayhoe*, 1978; *Huyakorn et al.*, 1984, 1986]. Recently, *Allen and Murphy* [1985, 1986] and *Celia et al.* [1987] used a mixed form of Richards equation to derive numerical solution algorithms. These algorithms used collocation approximations in space, with Celia and coworkers using an alternating-direction version of the collocation approximation. Allen and Murphy called their approximation a "quasi-Newton" method while Celia and coworkers referred to the method as a "modified Picard" method. Both demonstrated excellent mass balance in their numerical solutions. *Zarba* [1988] used the modified Picard iteration method with both finite difference and finite element approximations in space and demonstrated perfect mass balance.

Milly [1986] and *Celia et al.* [1987] pointed out mass balance problems in standard *h*-based numerical solutions. *Milly* [1985] also presented a mass-conserving solution procedure that used a modified definition of the capacity term to force global mass balance. He also pointed out the importance of mass lumping in finite element solutions to unsaturated flow. The significance of mass lumping has also been suggested by others, including *Neuman* [1973] and *Cooley* [1983]; *Huyakorn et al.* [1984, 1986] report good finite element solutions without mass lumping. Neuman also suggested that evaluation of the nonlinear coefficients at the $n + \frac{1}{2}$ time level, in the context of a fully implicit approximation, improved convergence in iteration. All of the references cited above used a one-step Euler time-marching algorithm, and most began with the *h*-based version of Richards equation.

This paper demonstrates the problems inherent in *h*-based, one-step Euler solutions. It also demonstrates the relationship between the modified Picard method and the standard *h*-based Picard method and shows that the modified Picard approach is to be preferred. Arguments are also presented that show why mass lumping should be used in finite element solutions.

APPROXIMATION METHODS

Standard Approximations

Let us begin by examining the standard fully implicit Picard method applied to the *h*-based equation (1). Temporal discretization of (1) using a backward Euler method may be written as

$$C^{n+1} \frac{H^{n+1} - H^n}{\Delta t} - \nabla \cdot K^{n+1} \nabla H^{n+1} - \frac{\partial K^{n+1}}{\partial z} = 0 \qquad (4)$$

where H^n denotes the approximate value of h at the nth discrete time level ($t = t^n$), $\Delta t \equiv t^{n+1} - t^n$ is the time step, C^{n+1} and K^{n+1} denote specific moisture capacity and hydraulic conductivity evaluated using H^{n+1}, respectively, and the solution is assumed to be known at time level n and unknown at time level $n + 1$. Because C and K are nonlinear functions of h, some linearization must be introduced into (4). The Picard iteration method involves sequential estimation of the unknown H^{n+1} using the latest estimates of C^{n+1}, K^{n+1}. If m identifies iteration level, then the Picard iteration scheme can be written as

$$C^{n+1,m} \frac{H^{n+1,m+1} - H^n}{\Delta t} - \nabla \cdot K^{n+1,m} \nabla H^{n+1,m+1} - \frac{\partial K^{n+1,m}}{\partial z} = 0 \qquad (5a)$$

This equation may be rewritten in the following equivalent form,

$$\begin{aligned} &C^{n+1,m} \frac{H^{n+1,m+1} - H^{n+1,m}}{\Delta t} \\ &\quad - \nabla \cdot K^{n+1,m} \nabla (H^{n+1,m+1} - H^{n+1,m}) \\ &= -C^{n+1,m} \frac{H^{n+1,m} - H^n}{\Delta t} + \nabla \cdot K^{n+1,m} \nabla H^{n+1,m} \\ &\quad + \frac{\partial K^{n+1,m}}{\partial z} \\ &\equiv R_P^{n+1,m} \end{aligned} \qquad (5b)$$

In (5*b*), the right side is a measure of the amount by which the temporally discretized equation fails to be satisfied by the *m*th iterative estimate $H^{n+1,m}$. This is defined as the residual associated with the Picard iteration, R_P. Upon convergence in iteration, both $R_P^{n+1,m}$ and the difference in iteration $(H^{n+1,m+1} - H^{n+1,m})$ approach zero.

To demonstrate the behavior of this approximation, let us consider the case of one spatial dimension (z). If subscript i denotes spatial location ($z_i = i(\Delta z)$ for constant spacing Δz), then the standard finite difference (in space) approximation to (5) is

$$
\begin{aligned}
&C_i^{n+1,m}\frac{\delta_i^m}{\Delta t} - \frac{1}{(\Delta z)^2}\\
&\quad\cdot[K_{i+1/2}^{n+1,m}(\delta_{i+1}^m - \delta_i^m) - K_{i-1/2}^{n+1,m}(\delta_i^m - \delta_{i-1}^m)]\\
&= \frac{1}{(\Delta z)^2}[K_{i+1/2}^{n+1,m}(H_{i+1}^{n+1,m} - H_i^{n+1,m})\\
&\quad - K_{i-1/2}^{n+1,m}(H_i^{n+1,m} - H_{i-1}^{n+1,m})]\\
&\quad + \frac{K_{i+1/2}^{n+1,m} - K_{i-1/2}^{n+1,m}}{\Delta z} - C_i^{n+1,m}\frac{H_i^{n+1,m} - H_i^n}{\Delta t}\\
&\equiv (R_i^{n+1,m})_{\mathrm{PFD}}
\end{aligned}
\tag{6}
$$

where the dependent variable is the increment in iteration $\delta_i^m \equiv (H_i^{n+1,m+1} - H_i^{n+1,m})$. The right side of (6) is now a measure of the amount by which the *m*th iterate fails to satisfy the finite difference approximation. The residual $(R_i^{n+1,m})_{\mathrm{PFD}}$ is an error measure for the finite difference spatial approximation coupled with the Picard iteration procedure.

The finite element approximation to (5*b*) is usually generated using piecewise linear basis functions to approximate H as well as the coefficients C and K, namely,

$$h(z, t) \simeq \sum_{j=0}^{E} H_j(t)\phi_j(z) \tag{7a}$$

$$K(h) \simeq \sum_{j=0}^{E} K(H_j)\phi_j(z) = \sum_{j=0}^{E} K_j\phi_j(z) \tag{7b}$$

$$C(h) \simeq \sum_{j=0}^{E} C(H_j)\phi_j(z) = \sum_{j=0}^{E} C_j\phi_j(z) \tag{7c}$$

where $\phi_j(z)$ is the standard piecewise linear basis function, E is the number of elements, and $N = E + 1$ is the number of nodes. Evaluation of the integrals that occur in the finite element formulation leads to the following discrete equation for the piecewise linear finite element approximation,

$$
\begin{aligned}
&\bar{C}_{i-1}\frac{\delta_{i-1}^m}{\Delta t} + \bar{C}_i\frac{\delta_i^m}{\Delta t} + \bar{C}_{i+1}\frac{\delta_{i+1}^m}{\Delta t} - \frac{1}{(\Delta z)^2}\\
&\quad\cdot[K_{i+1/2}^{n+1,m}(\delta_{i+1}^m - \delta_i^m) - K_{i-1/2}^{n+1,m}(\delta_i^m - \delta_{i-1}^m)] = \frac{1}{(\Delta z)^2}\\
&\quad\cdot[K_{i+1/2}^{n+1,m}(H_{i+1}^{n+1,m} - H_i^{n+1,m})\\
&\quad - K_{i-1/2}^{n+1,m}(H_i^{n+1,m} - H_{i-1}^{n+1,m})]\\
&\quad + \frac{K_{i+1}^{n+1,m} - K_{i-1}^{n+1,m}}{2(\Delta z)} - \bar{C}_{i-1}\frac{H_{i-1}^{n+1,m} - H_{i-1}^n}{\Delta t}\\
&\quad - \bar{C}_i\frac{H_i^{n+1,m} - H_i^n}{\Delta t} - \bar{C}_{i+1}\frac{H_{i+1}^{n+1,m} - H_{i+1}^n}{\Delta t}\\
&\equiv (R_i^{n+1,m})_{\mathrm{PFE}}
\end{aligned}
\tag{8}
$$

where

$$\bar{C}_{i-1} \equiv \tfrac{1}{12}(C_{i-1}^{n+1,m} + C_i^{n+1,m}) \tag{9a}$$

$$\bar{C}_i \equiv \tfrac{1}{12}(C_{i-1}^{n+1,m} + 6C_i^{n+1,m} + C_{i+1}^{n+1,m}) \tag{9b}$$

$$\bar{C}_{i+1} \equiv \tfrac{1}{12}(C_i^{n+1,m} + C_{i+1}^{n+1,m}) \tag{9c}$$

$$K_{i-1/2}^{n+1,m} \equiv \tfrac{1}{2}(K_{i-1}^{n+1,m} + K_i^{n+1,m}) \tag{9d}$$

$$K_{i+1/2}^{n+1,m} \equiv \tfrac{1}{2}(K_i^{n+1,m} + K_{i+1}^{n+1,m}) \tag{9e}$$

and the right side $(R_i^{n+1,m})_{\mathrm{PFE}}$ is the finite element residual, analogous to that for the finite difference approximation (6). Comparison of (6) and (8) indicates that: (1) the finite element approximation spatially distributes the time approximation to the three nodes z_{i-1}, z_i, and z_{i+1}, while the finite difference approximation evaluates the time derivative at z_i only; (2) if $K_{i\pm1/2}$ of (6) are defined as the arithmetic average between K_i and $K_{i\pm1}$, then the spatial derivatives in the finite difference and finite element approximations are identical. While the geometric mean for $K_{i\pm1/2}$ is usually preferable [*Haverkamp and Vauclin*, 1979; *Hornung and Messing*, 1983], the arithmetic mean is used herein to maintain equivalence in the approximations to the spatial derivatives in the finite difference and finite element expressions. Therefore any difference in numerical solutions must be due to the time derivative term.

Given a set of soil properties $\theta(h)$ and $K(h)$, (6) and (8) can be solved for an approximation to the pressure head $h(z, t)$, which in turn provides the moisture content $\theta(z, t)$. As a first example, consider the data reported by *Haverkamp et al.* [1977]:

$$\theta(h) = \frac{\alpha(\theta_s - \theta_r)}{\alpha + |h|^{\beta}} + \theta_r \tag{10a}$$

$$K(h) = K_s\frac{A}{A + |h|^{\gamma}} \tag{10b}$$

where $\alpha = 1.611 \times 10^6$, $\theta_s = 0.287$, $\theta_r = 0.075$, $\beta = 3.96$, $K_s = 0.00944$ cm/s, $A = 1.175 \times 10^6$, and $\gamma = 4.74$. These data were used to solve an example of infiltration into a soil column. The problem considered corresponds to one solved by *Haverkamp et al.* [1977], with initial condition $h(z, 0) = -61.5$ cm, boundary conditions $h(40\text{ cm}, t) = h_{\mathrm{top}} = -20.7$ cm and $h(0, t) = h_{\mathrm{bottom}} = -61.5$ cm (the vertical dimension is assumed positive upward). Several pressure head profiles from the solution of (6) and (8) are shown in Figure 1. These solutions all correspond to the simulated solution at an elapsed

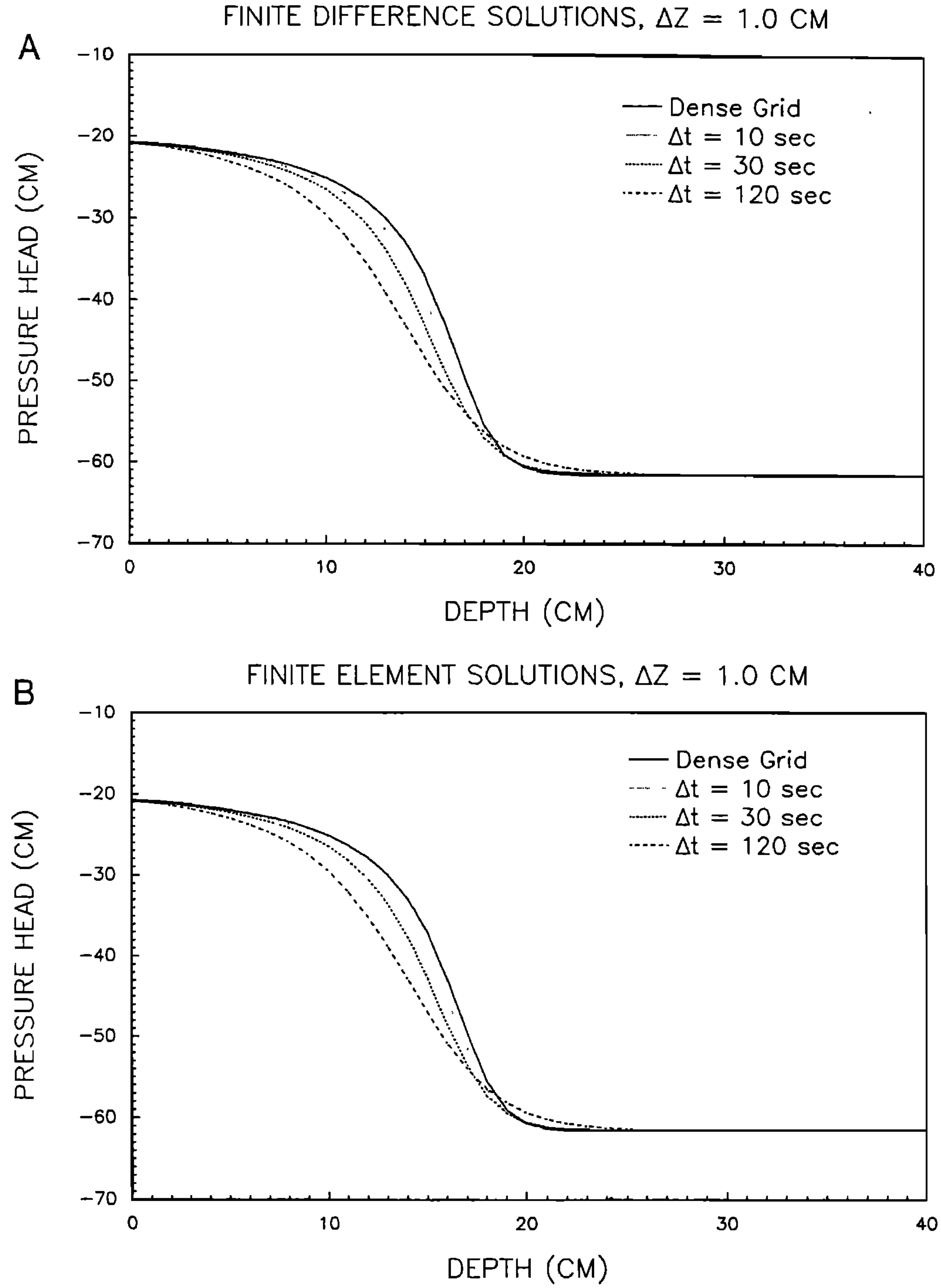

Fig. 1. (*a*) Finite difference and (*b*) finite element solutions using *h*-based equations with data of (10).

time of 360 s, using a node spacing of Δz of 1 cm and several different values of the time step Δt. The correct solution, based on solutions using very fine space and time steps, is also shown in the figure. It is clear that as Δt increases, the error in the solution increases, although the solutions are still convergent in iteration. For a time step of $\Delta t = 120$ s, the infiltration depth is underestimated by more than 20%.

One measure of a numerical simulator is its ability to conserve global mass over the domain of interest. Adequate conservation of global mass is a necessary but not sufficient condition for acceptability of a numerical simulator. To measure the ability of the simulator to conserve mass, let a mass balance measure be defined as follows:

$$\mathrm{MB}(t) \equiv \frac{\text{total additional mass in the domain}}{\text{total net flux into the domain}} \qquad (11)$$

where the additional mass is measured with respect to the initial mass in the system. For the finite difference approximation with first type boundary conditions, this is calculated as

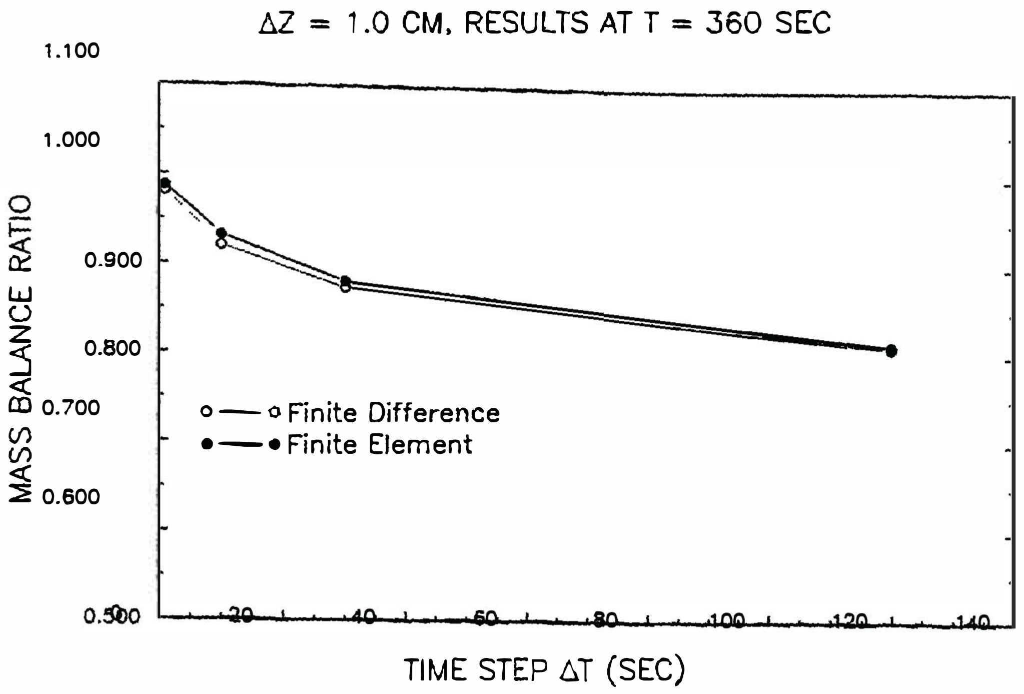

Fig. 2. Mass balance results for data of (10).

$$\mathrm{MB}_{\mathrm{FD}}(t^{n+1}) = \frac{\displaystyle\sum_{i=1}^{E-1} (\theta_i^{n+1} - \theta_i^0)(\Delta z)}{\displaystyle\sum_{j=1}^{n+1} \left\{ K_{N-1/2}^j \left[\frac{H_N^j - H_{N-1}^j}{\Delta z} + 1 \right] - K_{1/2}^j \left[\frac{H_1^j - H_0^j}{\Delta z} + 1 \right] \right\} (\Delta t)} \tag{12a}$$

where there are $N = E + 1$ nodes $\{z_0, z_1, \cdot\ \cdot, z_E\}$, and constant nodal spacing Δz is assumed. The finite element mass balance is of the form

$$\mathrm{MB}_{\mathrm{FEM}}(t^{n+1}) = \left[\sum_{i=1}^{E-1} [(\theta_i^{n+1} - \theta_i^0)(\Delta z)] + (\theta_0^{n+1} - \theta_0^0) \left(\frac{\Delta z}{2} \right) + (\theta_E^{n+1} - \theta_E^0) \left(\frac{\Delta z}{2} \right) \right] \cdot \left(\sum_{j=1}^{n+1} [(q_0^j - q_N^j)(\Delta t)] \right)^{-1} \tag{12b}$$

with q_0 and q_N being boundary fluxes calculated from the finite element equations associated with the boundary nodes z_0 and z_N.

Figure 2 shows mass balance as a function of Δt for the solutions of Figure 1. While there is a large range of time steps for which the solution converges in iteration, for most of this range the mass balance error is greater than 10%. Only for very small time steps does the solution exhibit mass balance errors of less than 5%. This behavior is observed for both finite difference and finite element approximations, with mass balance being slightly better for the finite element solutions. Similar observations on mass balance behavior for h-based formulations have been reported by *Milly* [1985].

Other material properties have been used in simulations and the behavior was generally analogous to that described above. One of these data sets was taken from D. Polmann (personal communication, 1988) and is an approximate description of a field site in New Mexico. The relevant material properties are

$$\theta(h) = \frac{\theta_s - \theta_r}{[1 + (\alpha|h|)^n]^m} + \theta_r \tag{13a}$$

$$K(h) = K_s \frac{\{1 - (\alpha|h|)^{n-1}[1 + (\alpha|h|)^n]^{-m}\}^2}{[1 + (\alpha|h|)^n]^{m/2}} \tag{13b}$$

where $\alpha = 0.0335$, $\theta_s = 0.368$, $\theta_r = 0.102$, $n = 2$, $m = 0.5$, $K_s = 0.00922$ cm/s. Initial and boundary conditions were taken as $h(z, 0) = -1000$ cm, $h(0, t) = h_{\mathrm{bottom}} = -1000$ cm, $h(60\text{ cm}, t) = -75$ cm. Figures 3 and 4 present solution profiles and mass balance results, respectively, for the data of equations (13). These figures show that for these data there is a much more pronounced difference between the finite difference and finite element solutions. The finite element solution has significantly larger mass balance error, underpredicts the infiltration depth to a greater extent than the finite difference solutions, and exhibits undershoot errors ahead of the infiltration front. Because the spatial distribution of the time derivative term in the finite element approximation is the only difference in the approximation, it must be the cause of these differences between the finite difference and finite element solutions. The comparison of finite difference and finite element solutions is discussed in more detail below.

The results presented above hold for any iterative solution

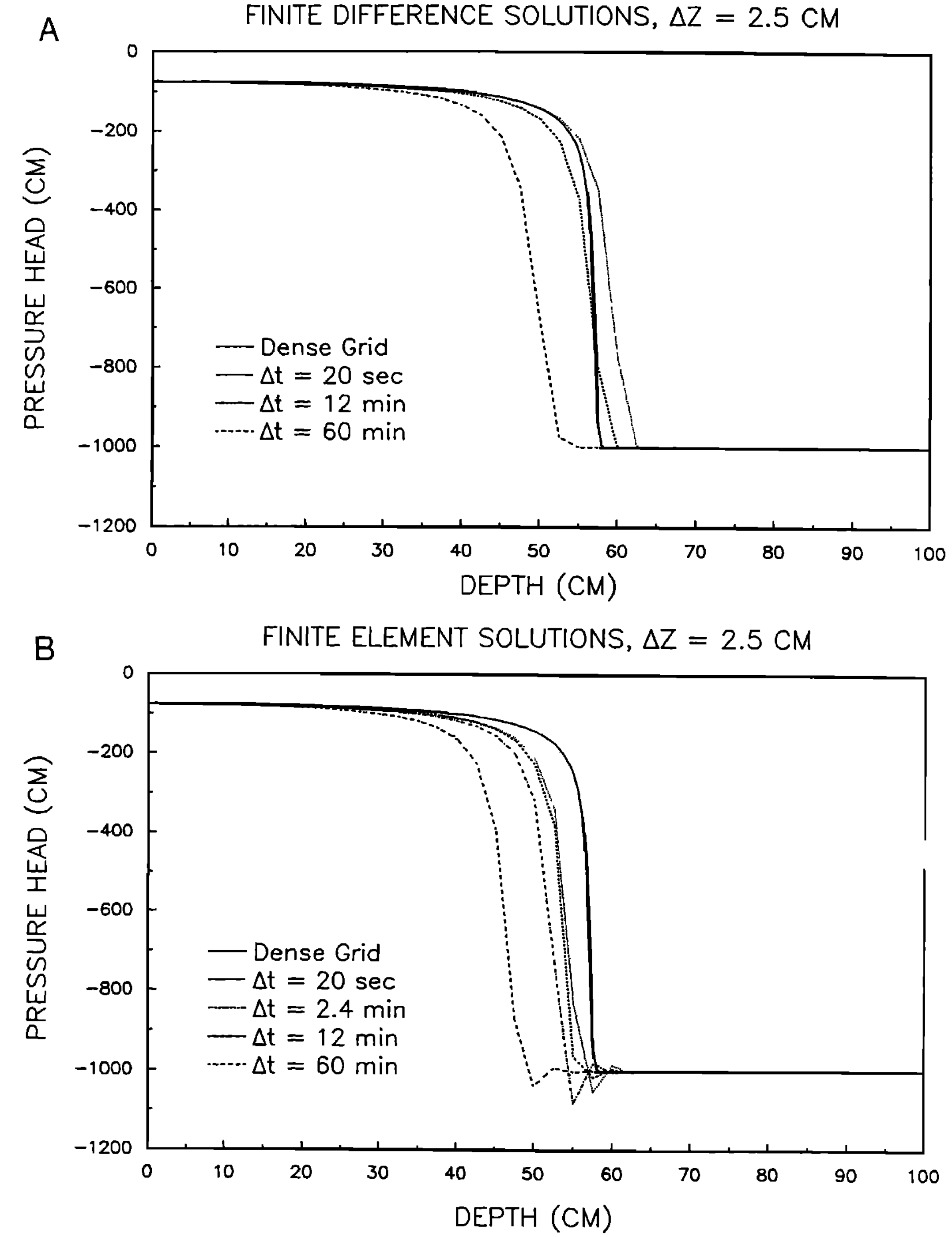

Fig. 3. (*a*) Finite difference and (*b*) finite element solutions using *h*-based equations with data of (13). Finite difference solution using Δt = 2.4 min did not converge in nonlinear iteration.

to the discretized version of the *h*-based (4), including Newton methods, because all convergent solutions should converge to the same solution. The only effect that different iteration methods have, once the discrete equations have been formed, is on convergence rates and computational efficiency. If both the Picard and Newton-Raphson iterative solutions for (6) converge, then they will converge to the same solution and therefore exhibit the same numerical performance. Thus mass balance or other numerical errors cannot be overcome by choosing a different iteration scheme. It is the original discretization that must be altered.

A Simple Mass-Conservative Scheme

One of the advantages of the θ-based equation is that discrete approximations to it, such as the finite element and finite difference methods used above, can be formulated so that they are perfectly mass conservative. However, because this form of the Richards equation degenerates in fully saturated media, and because material discontinuities produce discontinuous θ profiles, the θ-based equation is usually not used for general groundwater hydrology problems. Upon examination of the *h*-based equation and its numerical ap-

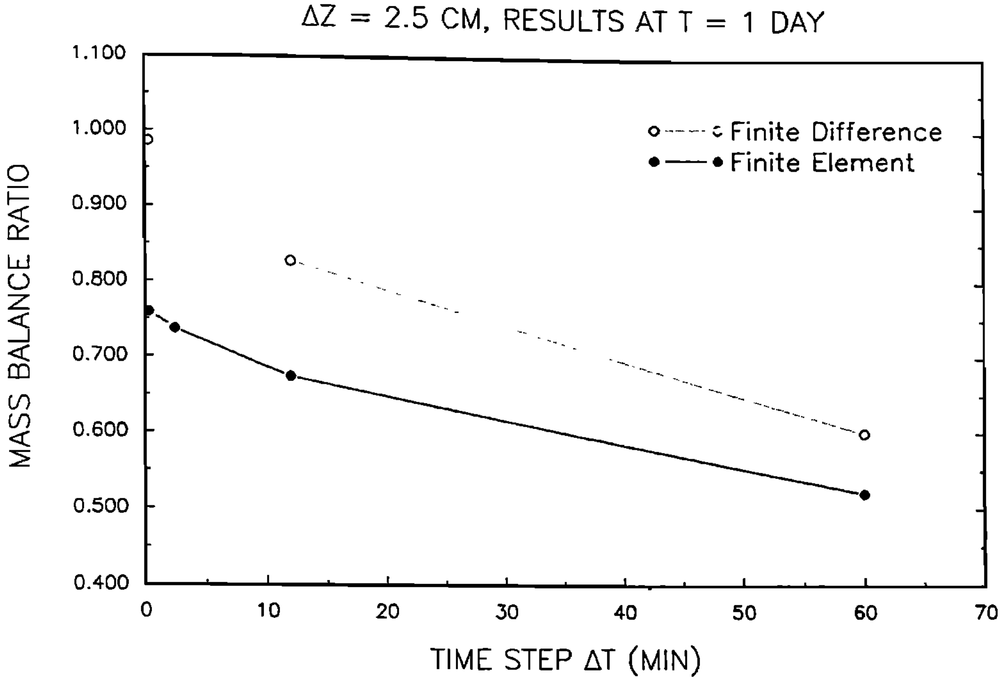

Fig. 4. Mass balance results for data of (13).

proximations, it can be seen that the reason for poor mass balance resides in the time derivative term. While $\partial\theta/\partial t$ and $C(\partial h/\partial t)$ are mathematically equivalent in the continuous partial differential equation, their discrete analogs are not equivalent. This inequality in the discrete forms is exacerbated by the highly nonlinear nature of the specific capacity term $C(h)$. This leads to significant mass balance errors in the h-based formulations because the change in mass in the system is calculated using discrete values of $\partial\theta/\partial t$ while the approximating equations use the expansion $C(h)$ $(\partial h/\partial t)$. In addition, the spatially distributed form that arises in the finite element approximation appears to further exacerbate the problems inherent in this term.

These observations lead to the conjecture that the mixed form of Richards equation (3) might be used to maintain the conservative property inherent in the θ-based equation, while providing the solution in terms of pressure head h. Such a solution would be mass conservative while avoiding the disadvantages of the θ-based form. The following approximation achieves this objective without introducing any additional computational burdens. It has been used previously, in conjunction with a collocation approximation in space, by *Allen and Murphy* [1985, 1986], *Celia and Pinder* [1986], and *Celia et al.* [1987].

To develop the discrete approximation based on the mixed form of Richards equation, begin with a backward Euler approximation in time coupled with a simple Picard iteration scheme. Let superscripts n and m again denote time level and iteration level, respectively. The backward Euler approximation is then written as

$$\frac{\theta^{n+1,m+1}-\theta^{n}}{\Delta t}-\frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial H^{n+1,m+1}}{\partial z}\right)-\frac{\partial K^{n+1,m}}{\partial z}=0 \qquad (14)$$

where superscripts indicate time and iteration levels. The key to the method is expansion of $\theta^{n+1,m+1}$ in a truncated Taylor series with respect to H, about the expansion point $H^{n+1,m}$, namely,

$$\theta^{n+1,m+1}=\theta^{n+1,m}+\left.\frac{d\theta}{dh}\right|^{n+1,m}(H^{n+1,m+1}-H^{n+1,m})+0(\delta^{2}) \qquad (15)$$

If all terms higher than linear are neglected in (15), and this equation is substituted into (14), there results

$$\left(\frac{1}{\Delta t}C^{n+1,m}\right)\delta^{m}+\frac{\theta^{n+1,m}-\theta^{n}}{\Delta t}-\frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial H^{n+1,m+1}}{\partial z}\right)-\frac{\partial K^{n+1,m}}{\partial z}=0$$

This equation can be rewritten in terms of the increment in iteration $\delta^{m}\equiv H^{n+1,m+1}-H^{n+1,m}$,

$$\left(\frac{1}{\Delta t}C^{n+1,m}\right)\delta^{m}-\frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial\delta^{m}}{\partial z}\right)=\frac{\partial}{\partial z}\left(K^{n+1,m}\frac{\partial H^{n+1,m}}{\partial z}\right)+\frac{\partial K^{n+1,m}}{\Delta z}-\frac{\theta^{n+1,m}-\theta^{n}}{\Delta t} \qquad (16)$$

Equation (16) is a general mixed-form Picard iteration, called a modified Picard approximation by *Celia et al.*, [1987] and by *Zarba* [1988]. Finite difference, finite element, or any other spatial approximation to (16) produces the final discrete form of the approximation. For example, the standard finite difference approximation in space leads to

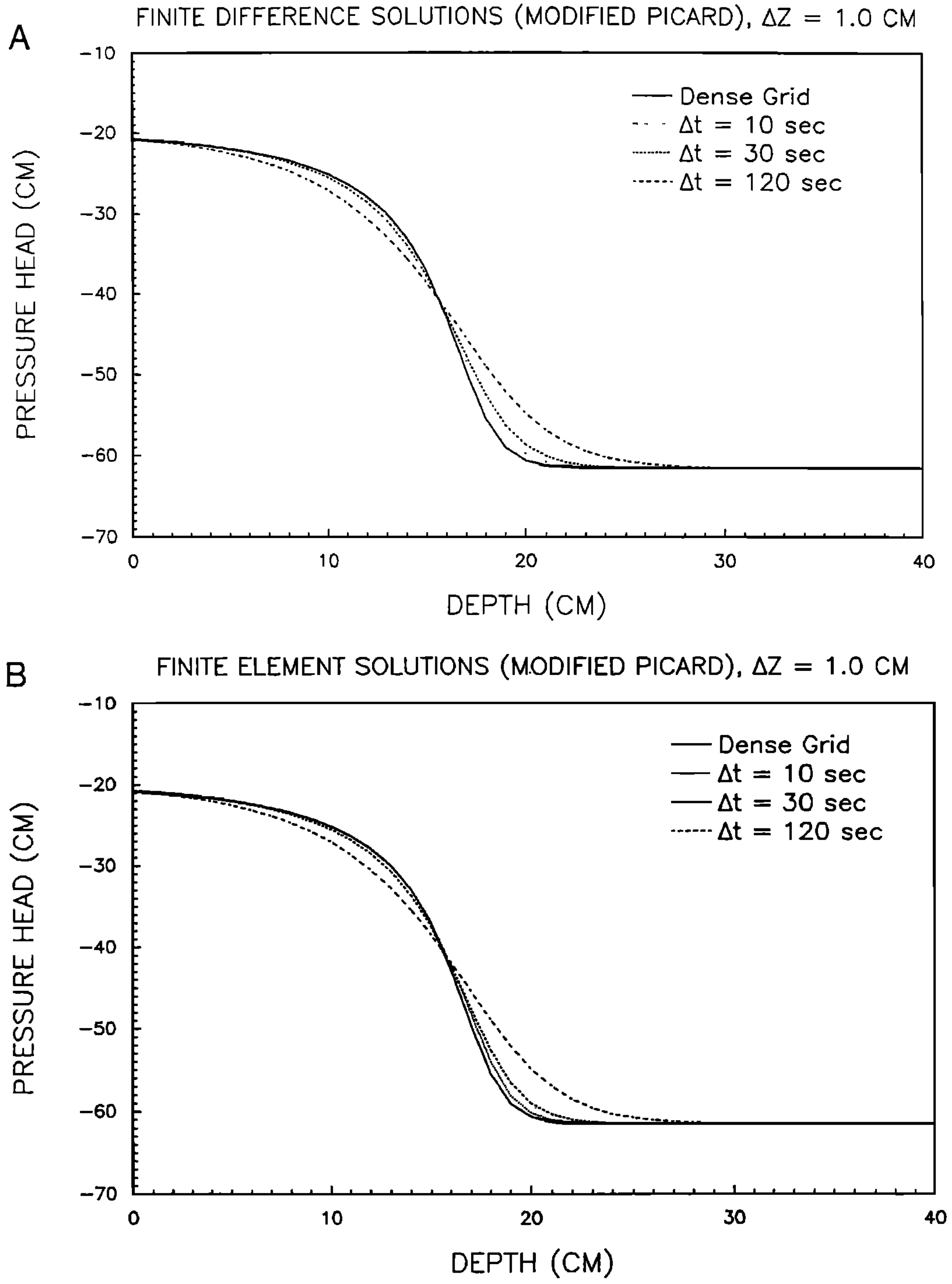

Fig. 5. (*a*) Finite difference and (*b*) finite element solutions using modified Picard method with data of (10).

$$C_i^{n+1,m}\frac{\delta_i^m}{\Delta t}-\frac{1}{(\Delta z)^2}\cdot[K_{i+1/2}^{n+1,m}(\delta_{i+1}^m-\delta_i^m)-K_{i-1/2}^{n+1,m}(\delta_i^m-\delta_{i-1}^m)]$$

$$=\frac{1}{(\Delta z)^2}[K_{i+1/2}^{n+1,m}(H_{i+1}^{n+1,m}-H_i^{n+1,m})-K_{i-1/2}^{n+1,m}(H_i^{n+1,m}-H_{i-1}^{n+1,m})]$$

$$+\frac{K_{i+1/2}^{n+1,m}-K_{i-1/2}^{n+1,m}}{\Delta z}-\frac{\theta_i^{n+1,m}-\theta_i^n}{\Delta t}$$

$$\equiv (R_i^{n+1,m})_{\text{MPFD}} \qquad (17)$$

Notice that the finite difference equation (17) is identical to that for the *h*-based form (equation (6)) except for the last term on the right side, which corresponds to the time derivative using the *m*th iteration level. It is precisely this

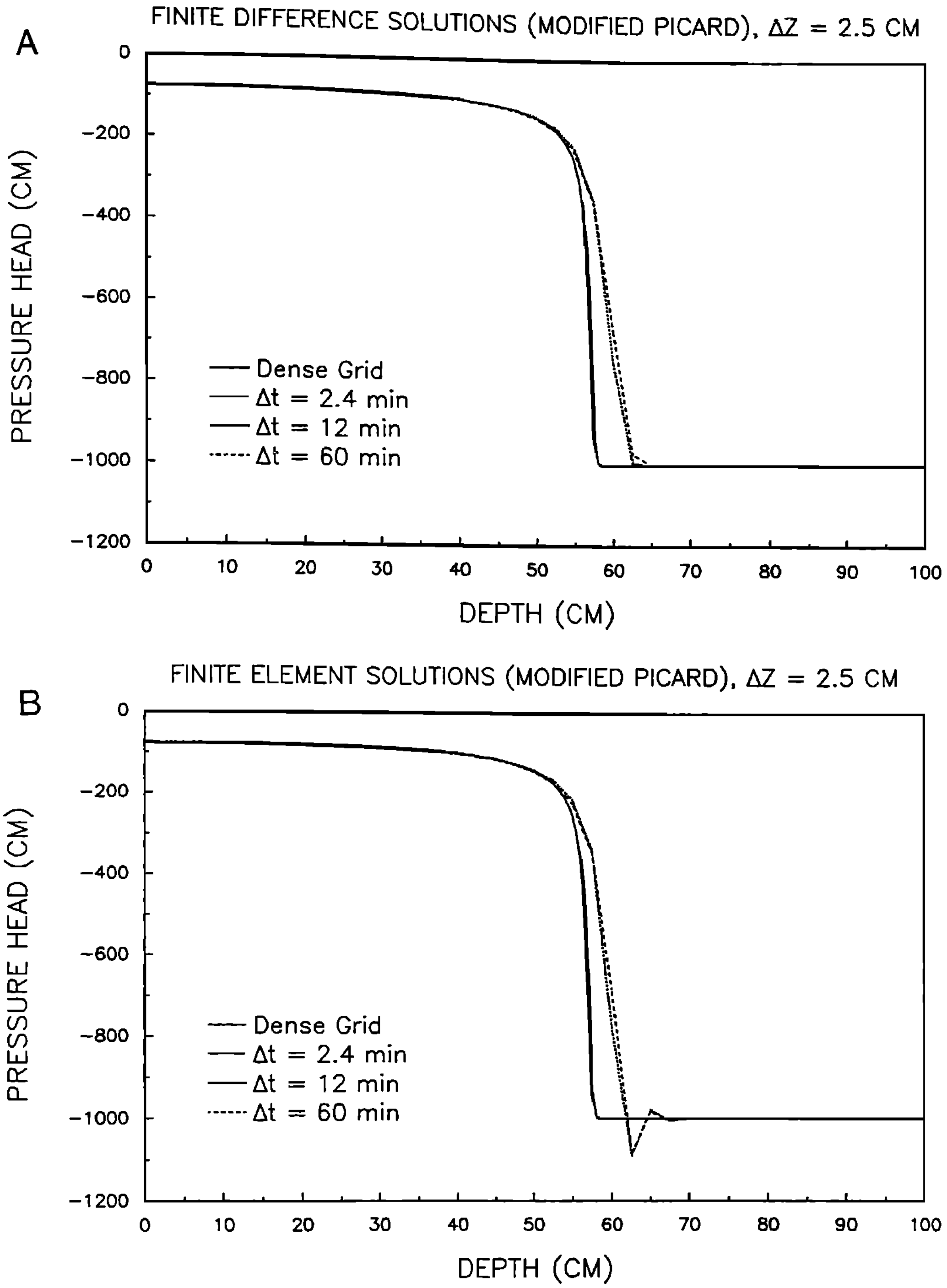

Fig. 6. (*a*) Finite difference and (*b*) finite element solutions using modified Picard method with data of (13).

difference that produces perfect mass balance. This is easily demonstrated by summing all finite difference equations (all nodes *i*); the only terms that do not cancel are the boundary fluxes and the change in total mass over the time step Δt. Therefore this set of equations possesses the conservative property [*Roache*, 1982], and global mass conservation is guaranteed. The finite element discretization can also be shown to possess the conservative property. In fact, the conservative property is maintained for all spatial approximations that maintain spatial symmetry. It also holds for all types of boundary conditions. However, it is generally restricted to one-step time-marching algorithms. Notice also that because the left sides of (6) and (17) are identical, the computational effort in solving the *h*-based and mixed forms are identical.

The influence of maintaining mass balance is illustrated in Figure 5. Here the same problem that was solved to generate the solutions shown in Figure 1 is solved using (17). While the errors associated with progressively larger time steps Δt in the *h*-based approximation led to poor mass balance and a progressively larger underprediction of infiltration depth, the mixed form maintains the correct infiltration depth but the moisture front is more diffuse as Δt increases. Imposition of global mass conservation forces the infiltration front to

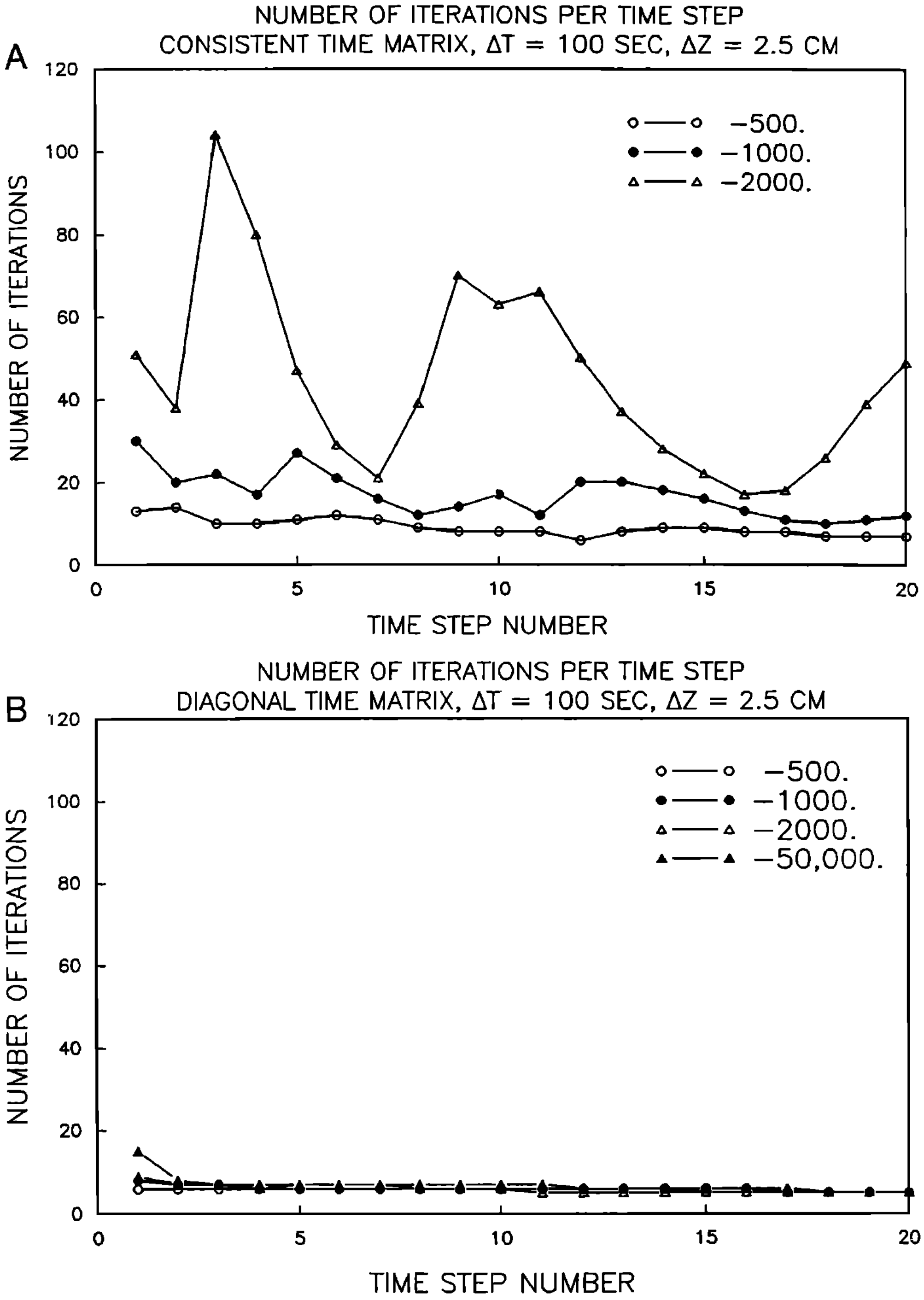

Fig. 7. (*a*) Number of iterations required for different initial conditions using modified Picard method with finite elements (consistent time matrix), (*b*) number of iterations required for different initial conditions using modified Picard method with finite differences (lumped finite elements), (*c*) typical solution using consistent time matrix, initial condition equal to −2000 cm, (*d*) typical solution using lumped time matrix, initial condition equal to −2000 cm, and (*e*) maximum undershoot for finite elements using consistent time matrices.

maintain its correct position, and the larger discretization errors appear as increased diffusion. No plot of mass balance as a function of time step is provided because the mass balance ratio of (11) is always unity.

Figure 6 shows numerical solutions using the mixed formulation for the data of D. Polmann (personal communication, 1988), which was used to produce the solutions in Figures 3 and 4. While the solutions of Figure 3 exhibited strong dependence on Δt, with deterioration of accuracy as Δt increased, the solutions of Figure 6 are insensitive to Δt. Thus in this case the mass conservative scheme serves to greatly reduce the influence of the time truncation errors, with the solutions of Figure 6 dominated by spatial error. In addition, whereas the h-based formulation exhibited differences in the location of the infiltration front between the finite element and finite difference solutions, these differ-

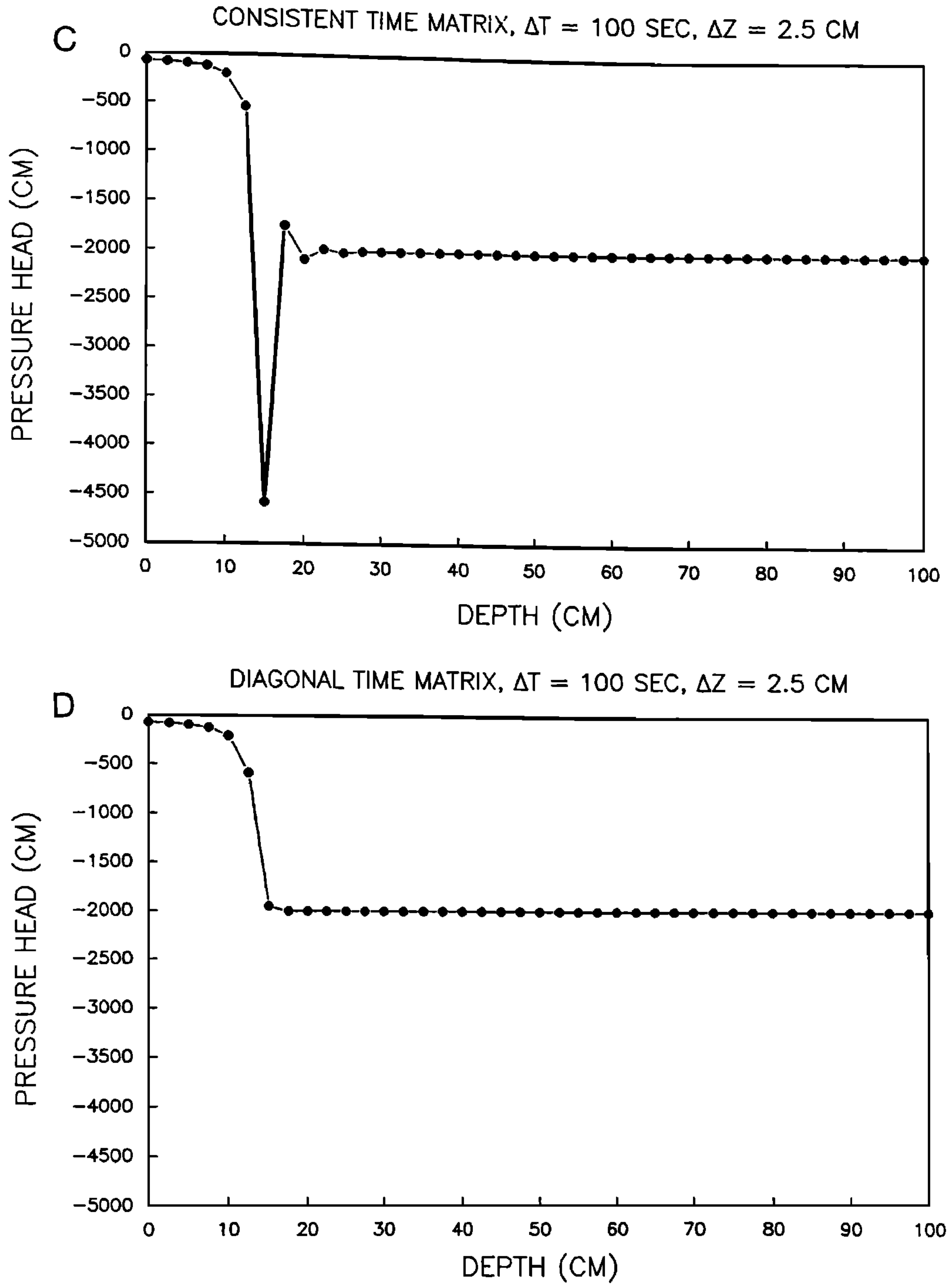

Fig. 7. (continued)

ences are essentially eliminated by use of the mixed formulation.

However, the finite element solutions still exhibit significant leading oscillations ahead of the moisture front. Such oscillations are not present in the finite difference approximation. Therefore while the modified linearization based on the mixed form of Richards equation improves both the finite element and finite difference solutions, it does not solve all problems associated with the finite element approximation. This demonstrates that attainment of proper mass balance is necessary but not sufficient to guarantee good numerical solutions. It also points to the importance of mass lumping in finite element solutions of unsaturated flow.

Finite Differences Versus Finite Elements: The Case for Mass Lumping

The previous numerical results demonstrate that the mixed formulation produces consistently superior numerical results, compared to analogous solutions based on the standard h-based methods. In addition, the results indicate that the finite element approximations may suffer from oscilla-

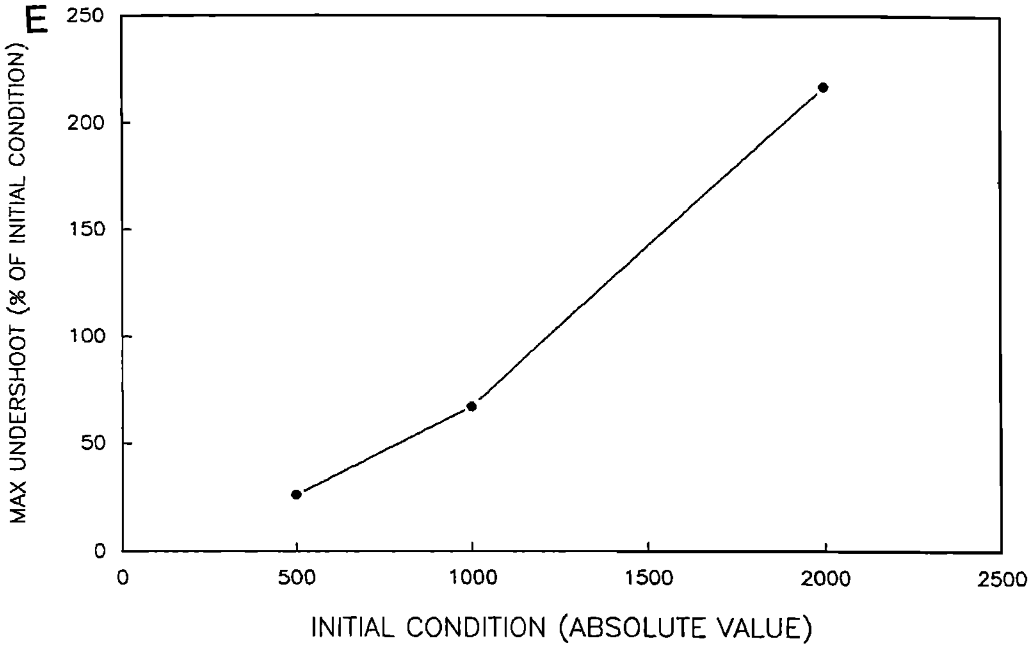

Fig. 7. (continued)

tory solutions. Such oscillations are not present in any finite difference solutions. Because the only difference between the two solution procedures is the treatment of the time derivative term, these results imply that diagonalized time (mass) matrices are to be preferred. Thus the unsaturated flow equation is one that benefits from mass lumping in finite element approximation.

The effect of diagonalization of the mass matrix is most evident in modeling infiltration into dry soils. To demonstrate this point, consider solution of the infiltration problem corresponding to (13) and the parameters that follow (13). To generate the solutions reported in Figures 3, 4, and 6, an initial condition of $h_0 = -1000$ cm was used. Now let the initial condition be varied from a maximum of −500 cm to a minimum of −50,000 cm. Results have been computed using the mixed formulation with both finite element and finite difference approximations in space. Salient features of these results are shown in Figure 7. In the first part of this figure, the number of iterations required for convergence is plotted as a function of time step for approximations that use Δz = 2.5 cm, Δt = 100 s, and an error tolerance on the residual of 10^{-8}. The finite difference (or "lumped" finite element) solutions converge rapidly, with the number of iterations increasing slightly over this range of initial conditions. In addition, all solutions are smooth, with maximum and minimum values corresponding to the top boundary condition and the initial condition, respectively. Conversely, the number of iterations for the finite element solutions increases dramatically as the initial condition becomes drier. No convergence was obtained for initial conditions of −5000 cm or drier. In addition, as the initial conditions became more negative, the finite element solutions generated progressively larger undershoots ahead of the pressure front. For h_0 = −500 cm, the minimum value over the first 40 time steps was −632 cm (a 26% undershoot); for h_0 = −2000 cm, the minimum value is −6337 cm (a 217% undershoot). Figures 7c–7e illustrate these results.

These results are illustrative of our numerical comparisons and indicate clearly that diagonal time matrices are superior to distributed matrices. This numerical evidence is complemented and explained by the fact that only when the time matrix is diagonal can the solution be guaranteed to be monotonic. This is because only in the diagonal case does the numerical solution satisfy the maximum principle [*Bouloutas*, 1989]. This principle states that the maximum value of the numerical solution is dictated by either the boundary conditions or the initial data. The implication is that the maximum value at any time t^{n+1} is less than or equal to the maximum value at the previous time level t^n. Similarly, the minimum value at t^{n+1} is greater than or equal to the minimum value at t^n. Thus oscillations in the solution cannot occur. This principle cannot be guaranteed to hold when the time matrix is not diagonal. Therefore finite element solutions admit oscillatory solutions. This argument about diagonal time matrices is true for any diffusion-type of equation, including the model linear diffusion equation. However, the highly nonlinear nature of the unsaturated flow equation appears to magnify the problem, leading the results such as those shown in Figure 7.

All of this leads to the conclusion that if finite element approximations are used to model unsaturated flow, they should be used in conjunction with a lumping procedure for the time matrix.

DISCUSSION

Based on the results presented above as well as many other numerical tests [*Zarba*, 1988], it appears that a mass-conserving approximation based on the mixed form of Richards equation, coupled with a lumped form of the time matrix, yields consistently reliable and robust numerical solutions for unsaturated flow problems. The *h*-based form of the equation, when used with a simple backward Euler approximation in time, gives consistently poor results, especially when a consistent finite element approximation is used. Such approximations should be avoided.

These observations are supported by other results as well. *Milly* [1985] reported poor mass balance for standard h-based simulators. *Celia et al.* [1987] reported similar mass balance problems when using a standard h-based simulator. In addition, *Celia et al.* [1987] used the mixed form of Richards equation in conjunction with a collocation approximation in space. Mass balance errors were consistently less than 1%. This indicates that the mixed formulation is robust with respect to mass balance, because the collocation approximation does not possess the conservative property [*Roache*, 1982]. The numerical method performs very well for more general unsaturated flow problems involving both infiltration and redistribution [*Bouloutas*, 1989]. Thus the procedure is not restricted in any way to only monotonic infiltration problems. Also, the conclusions presented herein do not change if the geometric mean, which is usually a better choice than the arithmetic mean, is used to define values of $K_{i\pm1/2}$.

The h-based formulation can be improved. As the development in this paper pointed out, it is the time derivative that is crucial to numerical performance. Our results indicate that a second-order correct, three-level approximation to the time derivative $\partial h/\partial t$,

$$\left.\frac{\partial h}{\partial t}\right|^{n+1} \simeq \frac{3h^{n+1} - 4h^n + h^{n-1}}{2(\Delta t)} \qquad (18)$$

can be used to produce significant improvements in the h-based methods [*Bouloutas*, 1989]. Other procedures such as evaluating the coefficients C and K in (5) at the $n + \frac{1}{2}$ time level [*Neuman*, 1973] are not nearly as helpful as is the use of (18). While the improved h-based solutions achieve better mass balance and improved accuracy, they still remain inferior to the mixed formulations. Thus the mass-conserving scheme presented above is to be preferred.

Finally, we have begun to extend these procedures to two- and three-dimensional unsaturated flow problems. Preliminary results indicate that the properties demonstrated herein for one-dimensional problems carry over to the multidimensional case. In particular, combination of the mass-conserving mixed formulation and mass lumping produces solutions that are consistently reliable and robust, even for very dry initial conditions.

CONCLUSIONS

Numerical solutions for the unsaturated flow equation can be substantially improved by use of a simple mass-conserving approximation. This approximation is based on the mixed form of Richards equation. It combines benefits inherent in both the θ-based and the h-based forms of the equation while circumventing major problems associated with each. These problems include poor mass balance and associated poor accuracy in h-based solutions, and restricted applicability of θ-based models. The method based on the mixed form of Richards equation is superior to h-based solutions while requiring no more computational effort. Because the problems associated with standard h-based solutions appear to be chronic, standard backward Euler approximations to the h-based form of the equation should be avoided. While Picard iteration was used herein, use of any other nonlinear solution technique, such as Newton-Raphson, will not alleviate these problems that are inherent in the h-based formulations.

The mass-conservative formulation is not sufficient to guarantee good approximations. In some situations, the spatial approximation used is also very important. Finite element approximations that use consistent formulations for the time matrix are generally inferior to finite difference (or lumped finite element) approximations. Results presented indicate that this becomes critical in solving problems of infiltration into initially very dry soils. The poor performance of consistent finite element solutions is due to loss of monotonicity, which results from a failure to satisfy a maximum principle. The appropriate solution methodology for general unsaturated flow problems is one that is based on the mixed form of Richards equation, and uses a lumped form of the time matrix.

Acknowledgments. This work was supported in part by the U.S. Nuclear Regulatory Commission under contract NRC-04-88-074, by the National Science Foundation under grant 8657419-CES, and by the U.S. Environmental Protection Agency under agreement CR-814946. Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency under assistance agreement CR-814946 to Princeton University, it has not been subjected to Agency review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

REFERENCES

Allen, M. B., and C. L. Murphy, A finite element collocation method for variably saturated flows in porous media, *Numer. Methods Partial Differential Equations*, *1*(3), 229–239, 1985.

Allen, M. B., and C. L. Murphy, A finite element collocation method for variable saturated flow in two space dimensions, *Water Resour. Res.*, *22*, 1537–1542, 1986.

Bouloutas, E. T., Improved numerical approximations for flow and transport in the unsaturated zone, Ph.D. thesis, Dep. of Civ. Eng., Mass. Inst. of Technol., Cambridge, 1989.

Celia, M. A., and G. F. Pinder, An alternating-direction collocation solution for the unsaturated flow equation, in *Proceedings of the VI International Conference on Finite Elements in Water Resources*, edited by S. da Costa et al., pp. 395–410, Computational Mechanics Publications, Southampton, England, 1986.

Celia, M. A., L. R. Ahuja, and G. F. Pinder, Orthogonal collocation and alternating-direction procedures for unsaturated flow problems, *Adv. Water Resour.*, *10*, 178–187, 1987.

Cooley, R. L., Some new procedures for numerical solution of variably saturated flow problems, *Water Resour. Res.*, *19*, 1271–1285, 1983.

Haverkamp, R., and M. Vauclin, A note on estimating finite difference interblock hydraulic conductivity values for transient unsaturated flow problems, *Water Resour. Res.*, *15*, 181–187, 1979.

Haverkamp, R., M. Vauclin, J. Touma, P. Wierenga, and G. Vachaud, Comparison of numerical simulation models for one-dimensional infiltration, *Soil Sci. Soc. Am. J.*, *41*, 285–294, 1977.

Hayhoe, H. N., Study of the relative efficiency of finite difference and Galerkin techniques for modeling soil-water transfer, *Water Resour. Res.*, *14*, 97–102, 1978.

Hillel, D., *Fundamentals of Soil Physics*, Academic, San Diego, Calif., 1980.

Hornung, U., and W. Messing, Truncation errors in the numerical solution of horizontal diffusion in saturated/unsaturated media, *Adv. Water Resour.*, *6*, 165–168, 1983.

Huyakorn, P. S., S. D. Thomas, and B. M. Thompson, Techniques for making finite elements competitive in modeling flow in variably saturated media, *Water Resour. Res.*, *20*, 1099–1115, 1984.

Huyakorn, P. S., E. P. Springer, V. Guvanasen, and T. D. Wadsworth, A three-dimensional finite-element model for simulating water flow in variably saturated porous media, *Water Resour. Res.*, *22*, 1790–1808, 1986.

Milly, P. C. D., A mass-conservative procedure for time-stepping in models of unsaturated flow, *Adv. Water Resour.*, *8*, 32–36, 1985.

Milly, P. C. D., Advances in modeling of water in the unsaturated zone, *Transp. Porous Media*, *3*, 491–514, 1988.

Narasimhan, T. N., and P. A. Witherspoon, An integrated finite difference method for analyzing fluid flow in porous media, *Water Resour. Res.*, *12*, 57–64, 1976.

Neuman, S. P., Saturated-unsaturated seepage by finite elements, *J. Hydraul. Div. Am. Soc. Civ. Eng.*, *99*(HY12), 2233–2250, 1973.

Nielsen, D. R., M. T. van Genuchten, and J. W. Biggar, Water flow and solute transport processes in the unsaturated zone, *Water Resour. Res.*, *22*, 89S–108S, 1986.

Roache, P. J., *Computational Fluid Dynamics*, Hermosa Publishers, Albuquerque, N. M., 1982.

Zarba, R. L., A numerical investigation of unsaturated flow, M.S. thesis, Dep. of Civ. Eng., Mass. Inst. of Technol., Cambridge, 1988.

E. T. Bouloutas and M. A. Celia, Water Resources Program, Department of Civil Engineering and Operations Research, Princeton University, Princeton, NJ 08544.

R. L. Zarba, Camp Dresser and McKee Inc., One Center Plaza, Boston, MA 02108.

(Received July 17, 1989;
revised November 27, 1989;
accepted December 29, 1989.)