Advances in Water Resources Vol. 22, No. 5, pp. 419–430, 1999
© 1999 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0309-1708/99/$ - see front matter

PII: S0309-1708(98)00032-3

Long-term simulation of water movement in soils using mass-conserving procedures

Peter Berg

Department of Environmental Sciences, University of Virginia, Charlottesville, VA 22903, USA

(Received 29 December 1997; revised 3 July 1998; accepted 1 August 1998)

Three implicit numerical procedures for solving Richards' equation were compared. In some tests non-linear source terms were included. The main focus is on two mass-conserving procedures where the highly non-linear nature of Richards' equation is overcome by substantially different iterative techniques. In the first procedure (P1), which is well described in the literature, the time step is set back when difficulties in obtaining a convergent solution occur and any source terms are included explicitly. The second procedure (P2), which is presented in this paper, uses a fixed time step and convergence is obtained through a gradually increasing underrelaxation technique. Source terms are included implicitly, and in addition, a linear dependency between source terms and the water potential is included directly in the solution. The third procedure (P3) is based on a standard implicit discretization of Richards' equation, which does not facilitate mass conservation. All tests showed that mass conservation can be obtained without the added cost of more complicated programming or a longer calculation time. In challenging tests where fluctuating boundary conditions were specified as known fluxes, the total number of iterations needed in P1 to simulate the tests was many orders of magnitude higher than that for P2. In tests including source terms, the more advanced treatment of these in P2 resulted in convergence rates that were significantly higher that for the explicit coupling in P1. The gradually increasing underrelaxation technique and the more advanced treatment of source terms in P2 have proven to be effective approaches, particularly in long-term simulations of water movements in soils where calculation time can be prohibitive. As a secondary benefit, the implementation of P2 is less complicated than P1. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: long-term simulations, Richards' equation, mass-conserving procedures.

1 INTRODUCTION

Although Richards' equation[18] is only strictly valid for incompressible soils where water flows isothermally in the soil matrix (no macro-pore flow), it is widely used to describe transient water flow in realistic field situations. As outlined by Parlange *et al.*,[16] analytical solutions of Richards' equation have become more advanced, but numerical techniques must still be applied in order to obtain a general solution when soil inhomogeneities, fluctuating boundary conditions, plant uptake, and drainage are taken into account. This has been done in many models such as DAISY,[6] SOIL/SOILN[10,12] and RZWQM[1] where transport of water is coupled to plant growth and the transport of energy and nitrogen in a detailed description of the soil–plant–atmosphere system. Such models are typically used in long-term studies where many years are simulated repeatedly and often require excessive calculation time. Because of the strong non-linear nature of Richards' equation, the numerical sub-solution of this equation is commonly the most challenging part, and many studies have focused on developing more effective and reliable numerical procedures for solving this particular equation.[13] As shown by Milly and Eagleson,[15] common standard discretization techniques give a finite mass balance error that cannot be neglected unless extremely small time steps are used. Thus, it was a major step forward when different mass-conserving schemes were presented in the 1980s that allowed much larger time steps, while still keeping a good discrete approximation of Richards' equation.[15,5,14,2,4,3]

Issues other than obtaining mass conservation when solving Richards' equation are clearly important as well; the strong non-linear nature of the Richards' equation often causes serious problems in the iterative search for a convergent solution. In extreme situations, some schemes simply fail to find a convergent solution. In addition, source terms that are non-linear functions of the soil water potential, which normally are present when dealing with realistic field situations, can have additional negative effects on a procedure's ability to find a convergent solution. These problems are addressed in the present study where the performance and ease of implementation for three procedures for solving Richards' equation are compared.

The first procedure (referred to as P1) is based on the popular implicit and mass-conserving iteration scheme by Celia *et al.*,[3,4] and is implemented as described by Ahuja *et al.*[1] The second procedure (P2) is a new procedure, where mass conservation is obtained by applying the approximation of the specific water content suggested independently by Cooley[5] and Milly.[14] The third procedure (P3) is based on a standard implicit discretization, and for that reason it does not facilitate mass conservation. The main focus is on the two mass-conserving procedures where the problems arising from the described non-linearities are solved in significantly different ways.

2 THEORY

All three procedures produce solutions of Richards' equation where a source term has been included:

$$\frac{\partial\theta}{\partial t}=\frac{\partial}{\partial z}\left(K\left(\frac{\partial\psi}{\partial z}-1\right)\right)+S \tag{1}$$

where θ is the volumetric soil water content, ψ is the soil water potential, K is the hydraulic conductivity, t is the time, z is the depth below the soil surface, assumed to be positive downward, and S is the source term.

The iteration schemes in the three procedures are listed below, where superscripts n and $n+1$ are used for the time levels at time t and $t+\Delta t$, respectively, and m and $m+1$ denote the iteration levels where values at level m are known.

2.1 Procedure P1

As described by Ahuja *et al.*,[1] an initial guess in each time step of the desired solution is calculated by a single Euler step using eqn (2):

$$\begin{aligned}\theta_j^{n+1,1}=\theta_j^n+\Delta t\Bigg(&\frac{K_{j+1/2}^n}{0.5\Delta z_j(\Delta z_j+\Delta z_{j+1})}(\psi_{j+1}^n-\psi_j^n)\\&-\frac{K_{j-1/2}^n}{0.5\Delta z_j(\Delta z_{j-1}+\Delta z_j)}(\psi_j^n-\psi_{j-1}^n)\\&-\frac{K_{j+1/2}^n-K_{j-1/2}^n}{\Delta z_j}\Bigg)\end{aligned} \tag{2}$$

Based on values of $\theta_j^{n+1,1}$, values of $\psi_j^{n+1,1}$ are calculated and the following iteration steps are carried out using the mass-conserving iteration scheme derived by Celia *et al.*:[3]

$$\begin{aligned}&\frac{\hat{C}_j^{n+1,m}}{\Delta t}\delta_j^{n+1,m+1}-\frac{K_{j+1/2}^{n+1,m}}{0.5\Delta z_j(\Delta z_j+\Delta z_{j+1})}\\&\quad\times(\delta_{j+1}^{n+1,m+1}-\delta_j^{n+1,m+1})\\&\quad+\frac{K_{j-1/2}^{n+1,m}}{0.5\Delta z_j(\Delta z_{j-1}+\Delta z_j)}(\delta_j^{n+1,m+1}-\delta_{j-1}^{n+1,m+1})\\&=\frac{K_{j+1/2}^{n+1,m}}{0.5z_j(\Delta z_j+\Delta z_{j+1})}(\psi_{j+1}^{n+1,m}-\psi_j^{n+1,m})\\&\quad-\frac{K_{j-1/2}^{n+1,m}}{0.5\Delta z_j(\Delta z_{j-1}+\Delta z_j)}(\psi_j^{n+1,m}-\psi_{j-1}^{n+1,m})\\&\quad-\frac{K_{j+1/2}^{n+1,m}-K_{j-1/2}^{n+1,m}}{\Delta z_j}-\frac{\theta_j^{n+1,m}-\theta_j^n}{\Delta t}+S_j^n\end{aligned} \tag{3}$$

where

$$\hat{C}_j^{n+1,m}=\left(\frac{d\theta}{d\psi}\right)^{n+1,m} \tag{4}$$

and

$$\delta_j^{n+1,m+1}=\psi_j^{n+1,m+1}-\psi_j^{n+1,m} \tag{5}$$

Since all values at iteration level m are known, eqn (3) describes a linear tridiagonal system of equations in $\delta_j^{n+1,m+1}$. After the system of equations is solved using the Thomas algorithm,[17] new values of $\psi_j^{n+1,m+1}$ are calculated using eqn (5) and the process is repeated (eqns (3)–(5)) until a convergent solution is found (the criterion is described below). If convergence fails (the number of iterations exceeds a certain maximum before the criterion for convergence is reached), the time step is cut in half and the entire process is repeated starting with eqn (2). If convergence still fails, the time step is set back again and the process is repeated. This setting back of the time step will continue until a convergent solution is found or a predefined minimum time step is reached. If a convergent solution is not found using this minimum time step, the initial guess given by the Euler step is used as the solution. After a solution is found, and if the time step is less than a predefined maximum, the time step is increased by a given factor for future calculations. It should be noted that $\delta_j^{n+1,m+1}$ (and the entire left side of eqn (3)), goes toward zero when a convergent solution is approached. The source term is included explicitly as it is evaluated at time level n (eqn (3)).

2.2 Procedure P2

The derivation of all equations used in this procedure is summarized in Appendix A. All iteration steps are done using the iteration scheme given by eqn (6) with a fixed time step. In the first iteration step, values of $\hat{C}_j^{n+1,1}$, $K_{j+1/2}^{n+1,1}$, $S_{pj}^{n+1,1}$, and $S_{cj}^{n+1,1}$ are taken from the last iteration in the previous time step.

$$\frac{\hat{C}_j^{n+1,m}}{\Delta t}\psi_j^{n+1,m+1} - \frac{K_{j+1/2}^{n+1,m}}{0.5\Delta z_j(\Delta z_j + \Delta z_{j+1})} \times (\psi_{j+1}^{n+1,m+1} - \psi_j^{n+1,m+1}) + \frac{K_{j-1/2}^{n+1,m}}{0.5\Delta z_j(\Delta z_{j-1} + \Delta z_j)}(\psi_j^{n+1,m+1} - \psi_{j-1}^{n+1,m+1}) - S_{pj}^{n+1,m}\psi_j^{n+1,m+1} = \frac{\hat{C}_j^{n+1,m}}{\Delta t}\psi_j^n + \frac{K_{j-1/2}^{n+1,m} - K_{j+1/2}^{n+1,m}}{\Delta z_j} + S_{cj}^{n+1,m} \tag{6}$$

where

$$\hat{C}_j^{n+1,m} = \frac{\theta_j^{n+1,m} - \theta_j^n}{\psi_j^{n+1,m} - \psi_j^n} \tag{7}$$

and

$$S_{pj}^{n+1,m} = \left(\frac{\mathrm{d}S}{\mathrm{d}\psi}\right)_j^{n+1,m} \text{ and } S_{cj}^{n+1,m} = S_j^{n+1,m} - \left(\frac{\mathrm{d}S}{\mathrm{d}\psi}\right)_j^{n+1,m}\psi_j^{n+1,m} \tag{8}$$

Eqn (3) describes a linear tridiagonal system of equations in $\psi_j^{n+1,m+1}$ that is solved in each iteration step using the Thomas algorithm.[17] If a convergent solution is not found after a couple of iterations, the solution is underrelaxed using eqn (9):

$$\psi_j^{n+1,m+1,u} = (1-\alpha^m)\psi_j^{n+1,m+1} + \alpha^m\psi_j^{n+1,m} \tag{9}$$

where the relaxation factor α^m is gradually increased with the number of iterations used in the actual time step. Immediately after the underrelaxation is done, $\psi_j^{n+1,m+1}$ is given the value of $\psi_j^{n+1,m+1,u}$. The source term is included implicitly in the solution (eqn (6)), and a linear dependency between the source term and ψ are included directly as well. The necessary linearization of any source terms, given by eqn (8), is derived and discussed in detail in Appendix A. The scheme is mass-conserving because of the approximation of $\hat{C}_j^{n+1,m}$ given by eqn (7). This approximation was first proposed by Cooley[5] and Milly[14] using a finite element method, but it also arises naturally when using the control volume approach as shown in Appendix A.

2.3 Procedure P3

This procedure is obtained simply by substituting the approximation of $\hat{C}_j^{n+1,m}$ in P2 with eqn (4). Since no source terms are included in the tests involving P3, the source term is neglected. These changes lead to a non mass-conserving procedure that can be characterized as a standard implicit discretization scheme.

2.4 Boundary conditions

In all three procedures, a system of linear equations of the following form is solved in each iteration step

$$A_j^m\phi_{j-1}^{n+1,m+1} + B_j^m\phi_j^{n+1,m+1} + C_j^m\phi_{j+1}^{n+1,m+1} = D_j^m \tag{10}$$

where the unknown $\phi_j^{n+1,m+1}$ equals $\delta_j^{n+1,m+1}$ in P1 and $\psi_j^{n+1,m+1}$ in P2 and P3. The depth increments in the soil column, or the control volumes, are numbered from 0 to $N+1$, where control volume number 0 and $N+1$ are used for imposing the boundary conditions implicitly. The assignment of B_0^m, C_0^m and D_0^m as well as A_{N+1}^m, B_{N+1}^m and D_{N+1}^m for P2 and P3 is straightforward as shown in Appendix A. Imposing a known water potential as a boundary condition in P1 is also straightforward, but imposing a known flux is slightly more complicated. As outlined in Appendix A, an imposed flux q_t at the upper boundary can be written as

$$q_t = -K_{1/2}^{n+1,m}\left(\frac{\psi_1^{n+1,m+1} - \psi_0^{n+1,m+1}}{0.5\Delta z_1} - 1\right) \tag{11}$$

Substituting $\psi_1^{n+1,m+1}$ and $\psi_0^{n+1,m+1}$ with $\delta_1^{n+1,m+1} + \psi_1^{n+1,m}$ and $\delta_0^{n+1,m+1} + \psi_0^{n+1,m}$ (both derived from eqn (5)) leads to the following values for B_0^m, C_0^m and D_0^m

$$B_0^m = 1,\quad C_0^m = -1,\quad D_0^m = \psi_1^{n+1,m} - \psi_0^{n+1,m} - \frac{\Delta z_1}{2}\left(1 - \frac{q_t}{K_{1/2}^{n+1,m}}\right) \tag{12}$$

When the tridiagonal system of equations (eqn (10)) is solved, $\delta_0^{n+1,m+1}$ will be calculated so that the water flux at the soil surface equals q_t.

2.5 Criterion of convergence

As outlined in Appendix A, the maximum value of the following residual is used as a criterion for convergence in P2

$$R_j^{m+1} = |D_j^{m+1} - A_j^{m+1}\psi_{j-1}^{n+1,m+1} - B_j^{m+1}\psi_j^{n+1,m+1} - C_j^{m+1}\psi_{j+1}^{n+1,m+1}| \tag{13}$$

which expresses the mass balance error per unit time for control volume j. The coefficients A_j^{m+1}, B_j^{m+1}, C_j^{m+1} and D_j^{m+1} are evaluated after $\psi_j^{n+1,m+1}$ is found. This definition of a residual as an appropriate choice can be recognized by inserting A_j^{m+1}, B_j^{m+1}, C_j^{m+1} and D_j^{m+1} as they are defined in Appendix A into eqn (13) and then letting Δz_j go to zero, in which case R_j^{m+1} correctly expresses the difference between the fluxes toward and away from the point j. Exactly the same criterion for P1 is obtained by defining R_j^{m+1} as the right side of eqn (3) multiplied by Δz_j and evaluated at iteration level $m+1$.

When a boundary condition is specified as a known water potential the related residual, R_0^{m+1} or R_{N+1}^{m+1}, will automatically have a value of zero for all three procedures. When a known flux is specified, special precaution must be taken for P1. Inserting A_0^{m+1}, B_0^{m+1}, C_0^{m+1} and D_0^{m+1} as they are defined in Appendix A for the flux condition into eqn (13) gives the following residual for P2 and P3:

$$R_0^{m+1} = \left| -K_{1/2}^{n+1,m+1}\left(\frac{\psi_1^{n+1,m+1}-\psi_0^{n+1,m+1}}{0.5\Delta z_1}-1\right) - q_t \right| \tag{14}$$

which appropriately expresses the difference between the flux that actually is imposed and the flux that is desired. In order to obtain the same residual for P1, R_0^{m+1} must be defined as

$$R_0^{m+1} = \left| \frac{2K_{1/2}^{n+1,m+1}D_0^{m+1}}{\Delta z_1} \right| \tag{15}$$

where D_0^{m+1} is taken from eqn (12).

3 RESULTS AND DISCUSSION

One major problem when solving Richards' equation plus any source terms is the strong non-linearity of both the equation and the source terms. The most significant differences between the two mass-conserving procedures P1 and P2 are the way these problems are solved. In implicit procedures iterations are necessary and divergence in the iterative process is to be expected if no special action is taken. This problem is solved in P1 by setting back the time step and repeating the iteration process if convergence is not obtained within a certain number of iterations and, in extreme situations, when convergence is not obtainable, by using the explicit Euler step as a solution. In P2 the problem is solved rather simply but very effectively by applying the gradually increasing underrelaxation technique. The source term is included explicitly in P1 although the rest of the procedure is implicit. In P2 the source term is included implicitly, and furthermore, a linear dependency between the source term and ψ are also included directly. In addition to these differences, the way that mass conservation is achieved in P2 calls for fewer numerical operations than in P1, which can be recognized by comparing eqns (3) and (6). Finally, no initial guess is calculated in P2. As shown below, these differences result in significantly different performances of P1 and P2. It should be noted that if the source term is neglected and if the same time step is used, P1 and P2 will lead to exactly the same convergent solution. This can be seen by comparing eqns (3) and (6) and recalling that the left side in eqn (3) is going to zero when convergence is approached.

In terms of implementation, P1 is slightly more complicated to compute than P2. This is largely due to the possible interruption in P1 of an ongoing iteration process if convergence fails, followed by a new iteration process with a smaller time step. In addition, the Euler step must be implemented as a possible solution in P1 and used in extreme situations when a convergent solution cannot be found.

The three procedures have been compared in four tests where different boundary conditions, source terms and soil types were prescribed. The following specific values were used in all tests. In P1, 50 iterations were carried out before the time step was cut in half (see analysis in Test 3). The minimum time step size was defined as 0.001 times the normal marching (maximum) time step which is identical to the interval used by Ahuja *et al.*[1] After a solution was found with a time step that is smaller than this maximum, the time step was increased by a factor of 1.1 (see analysis in Test 3). The maximum time step was used as the one and only time step in P2 and P3. Furthermore, no underrelaxation was done in the first five iterations in P2 and P3, and after this a gradually increasing relaxation factor (α^m) was used, which varied linearly from 0.1 at five iterations to 0.9 at 100 iterations. The mass balance error (referred to as MBE below) was used in the evaluation of the three procedures (defined as *the total additional mass in the calculation domain* divided by *the time integrated flux over the soil surface* (in %)):

$$MBE = 100\left(\int_0^Z \theta\,\mathrm{d}z\bigg|_{t=0} - \int_0^Z \theta\,\mathrm{d}z\bigg|_{t=T} - \int_0^T K\left(\frac{\partial\psi}{\partial z}-1\right)\mathrm{d}t\bigg|_{z=0} + \int_0^T K\left(\frac{\partial\psi}{\partial z}-1\right)\mathrm{d}t\bigg|_{z=Z}\right)\bigg/\left(-\int_0^T K\left(\frac{\partial\psi}{\partial z}-1\right)\mathrm{d}t\bigg|_{z=0}\right) \tag{16}$$

where Z refers to the depth of the calculation domain and T to the final time of the calculation. All calculations were done with a real number representation between 15 and 16 significant digits (double precision in FORTRAN).

3.1 Test 1

Comparisons were made with one of the analytical solutions of Richards' equation defined by Ross and Parlange:[19]

$$\theta = \left(\frac{A}{K_1}\left(1-\exp\frac{-nK_1(At-z)}{D_1}\right)\right)^{1/n} \tag{17}$$

which is based on the assumptions that $D = D_1\theta^n$, $K = K_1\theta^{n+1}$ and $\theta = \exp(K_1\psi/D_1)$, where D is the soil water diffusivity and D_1, K_1, A and n are constants. The solution, in which units are arbitrary as long as they are consistent, describes infiltration into a soil with a uniform water content as an initial condition. With the same dimensionless values of constants as used by Ross and Parlange[19] ($D_1 = 100$, $K_1 = 1$, $A = 2$ and $n = 3$), two soil water profiles were calculated at times equal to 5 and 10 and

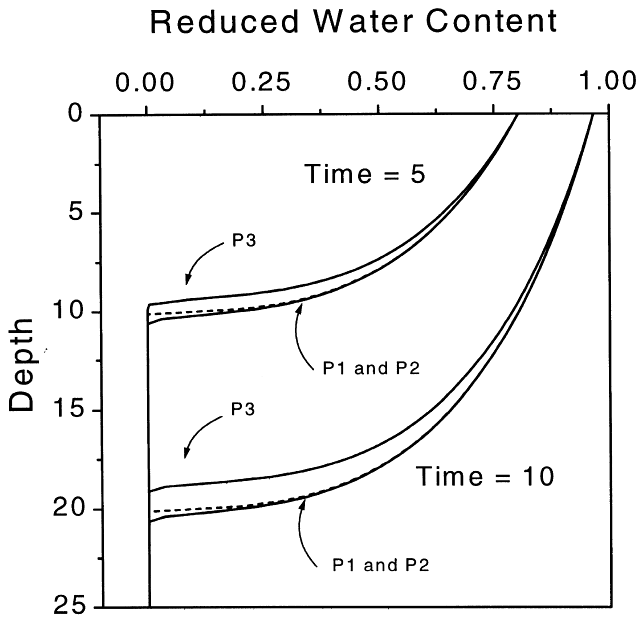

Fig. 1. Test 1: water content profiles for an infiltration event simulated with the upper boundary condition given as a known water potential. The solid lines represent simulated results of the three procedures and the dotted line represents an analytical solution by Ross and Parlange.[19]

compared with simulated profiles for all three procedures (Fig. 1). Known soil water potentials were specified as boundary conditions at the top ($z = 0$) and the bottom ($z = 25$) of the calculation domain. The maximum allowed time step was 0.05 and a resolution in depth of 0.25 was used. This resolution in time and space was found to be an appropriate choice for this problem. In additional simulations with P1 and P2 where the resolution in time and space was varied, a finer resolution gave almost the same results while a coarser resolution gave different results. The non mass-conserving procedure P3 gave a notably different result from the analytical solution (Fig. 1), while the identical profiles calculated with P1 and P2 agreed well with the analytical solution. The only visible deviation is seen near the depths where the soil water profile breaks. This was caused by an overestimation of the weighted hydraulic conductivity which was calculated as an arithmetic mean ($K_{j-1/2} = (K_{j-1} + K_j)/2$). Similar results were found by Warrick.[21] Additional simulations showed that an approximation of the weighted hydraulic conductivity as a geometric mean[7] ($K_{j-1/2} = (K_{j-1}K_j)^{1/2}$) will lead to unacceptable results due to a significant underestimation of the weighted hydraulic conductivity at these breaking points. It is important to note that the failure of the geometric mean is only due to the sudden break in the soil water profiles, and that the geometric mean is recommended as a general approximation in several studies.[7,20] No reduction in the time step was needed in P1 and the total numbers of iterations (summarized number of iterations over the entire simulation) were around 1000 for P1 and P2 and 2000 for P3. All profiles were calculated with a maximum allowed residual of 10^{-4}, which at time equal to 5 gave an MBE (eqn (16)) of 13% for P3 and less than 0.001% for both P1 and P2. Repeating the P3 simulation with a time step reduced by a factor of 5 still gave a considerable MBE of 3% and, of course, a large increase in the number of total iterations. A reduction in depth increments or a reduction of the maximum allowed residual had no effect on the MBE for P3. Although this is a very simple test, it shows clearly the considerable advantage of using mass-conserving procedures.

3.2 Test 2

The infiltration problem in Test 1 was simulated again, but this time with the upper boundary condition specified as a known flux. The first simulations of this test showed a strongly fluctuating MBE even for small variations of the maximum allowed residual ($R_{j\ \max}$). For that reason, the infiltration event was simulated repeatedly with all three procedures and with values of $R_{j\ \max}$ that were decreased in small steps from 1 to 10^{-7}. The results at time equal to 5 are shown in Fig. 2. Note first that the time integrated flux over the surface (Fig. 2A) was varying for values of $R_{j\ \max}$ larger than 10^{-3}, implying that $R_{j\ \max}$ must be specified to be less than 10^{-3} before the desired flux is imposed correctly.

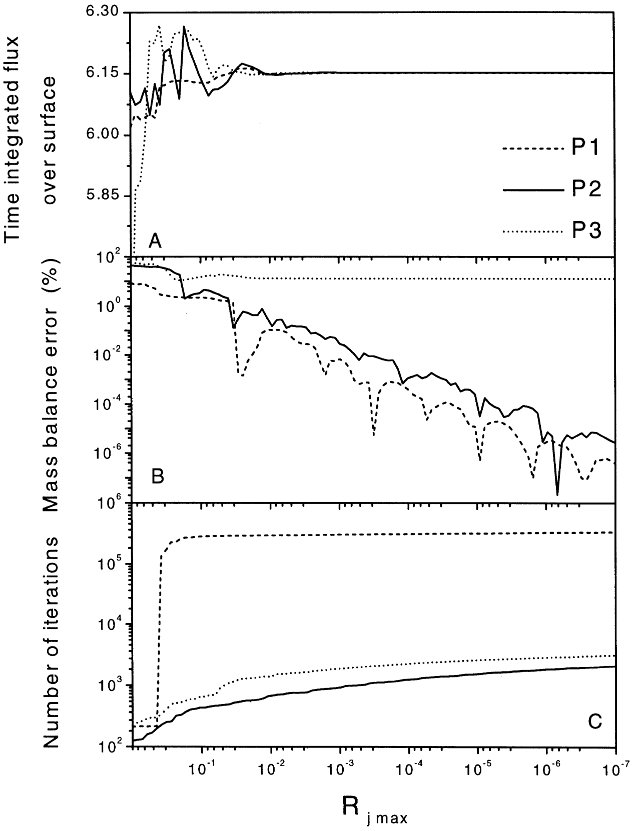

Fig. 2. Test 2: infiltration into the soil from Test 1, but now simulated with the upper boundary condition given as a known flux. Three characteristic parameters are shown as a function of the stop criterion in the iteration processes.

The MBE (Fig. 2B) for the non mass-conserving procedure P3 reached a minimum of 13% and stayed constant for decreasing values of $R_{j\ \max}$. The MBE for the two mass-conserving procedures P1 and P2 continued to decrease as expected for decreasing values of $R_{j\ \max}$. As mentioned above, pronounced fluctuations in MBE were seen for both P1 and P2, underlining the importance of using relatively small values of $R_{j\ \max}$ for both procedures. The number of total iterations used (Fig. 2C) in the two mass-conserving procedures is significantly different. On average, P1 used more that 300 times more iterations than P2 in the simulation of this infiltration event. Except for the largest values of $R_{j\ \max}$ (which gave an incorrectly imposed flux), the procedure P1 attempted to find a convergent solution in the very first part of the infiltration event operating with a time step of 0.00005 (the minimum allowed value). The initial guess given by the Euler step (eqn (2)) was used as the solution in almost all these time steps. Later in the infiltration event, the time step was gradually increased to 0.05 (the maximum value) for the rest of the simulation. This observed problem of P1 in obtaining convergence when steep gradients are present, has also been reported for the substantially identical scheme investigated by Johnsen *et al.*[11] Similar problems were not seen in the procedure P2.

3.3 Test 3

In a more demanding test, a layered soil column was constructed and simulated with highly dynamic boundary conditions for a period of one full year. The constructed soil column consisted of two layers of Yolo Light Clay[8] and one layer of sand.[8] The hydraulic properties for these two markedly different soil types are given by Haverkamp *et al.*[8] in the form of analytical expressions, and are shown in Fig. 3. The soil column and the separation into depth increments (control volumes), which increase steadily for depths greater than 44 cm, is shown in Fig. 4. The weighted hydraulic conductivity was calculated as the geometric mean ($K_{j-1/2} = (K_{j-1}K_j)^{1/2}$), as recommended by Haverkamp and Vauclin[7] and Schnabel and Richie.[20] As an upper boundary condition, measured values of precipitation with a time resolution of 3600 s were used in combination with a constant evaporation of 500 mm year^{-1}. The total precipitation equaled 548 mm year^{-1}, and the maximum value on a hourly, daily and weekly basis equaled 4 mm h^{-1}, 21 mm day^{-1} and 47 mm week^{-1}, respectively. As a lower boundary condition, a fixed soil water potential of 0 cm was used (i.e. a fixed water table). In P1, the time step was allowed to vary between 3.6 and 3600 s, which is equivalent to that used by Ahuja *et al.*,[1] and a fixed time step of 3600 s was used in P2 and P3. In a series of additional simulations with P1, different numbers of iterations were tried before reducing the time step. With values of this input parameter of 25, 50, 100 and 200, the following total number of iterations for simulating the full year was found to be 803 000, 697 000, 710 000 and 997 000 ($R_{j\ \max} =$ 10^{-12} m s^{-1}). Based on this, 50 iterations were chosen as an optimal number of iterations before the time step was set back. The optimal factor for increasing the time step in P1 was determined in a similar analysis. With values of this factor of 1.05, 1.1 to 1.2, the numbers of total iterations was calculated to be 829 000, 697 000 and 760 000 ($R_{j\ \max} =$ 10^{-12} m s^{-1}), based on which a factor of 1.1 was chosen. As in Test 2, the year was simulated repeatedly with varying values of $R_{j\ \max}$ and the results for all three procedures are shown in Fig. 5.

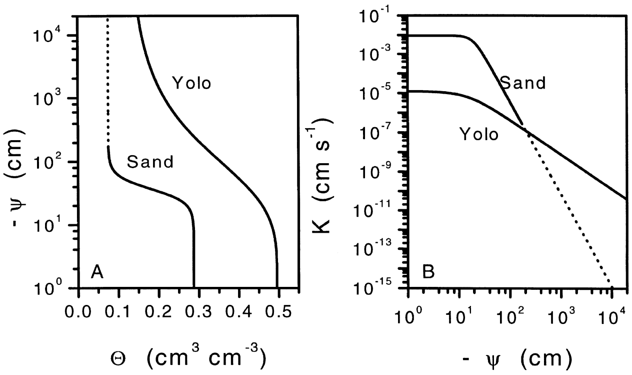

Fig. 3. Soil water content (A) and hydraulic conductivities (B) for the two markedly different soils used in Tests 3 and 4: the Yolo Light Clay and a sand soil (after Haverkamp *et al.*[8]).

The calculated soil water potential at the top of the upper clay layer varied between −3.8 and 0.71 m, while the variation in the top of the sand layer was between −0.90 and −0.50 m. Note first that a relatively small value of $R_{j\ \max}$ was needed before the desired net flux was imposed and an acceptable small MBE (eqn (16)) was achieved (Fig. 5A and 5B). The minimum MBE (Fig. 5B) for the non mass-conserving procedure P3 is high, 530%, corresponding to 250 mm. Achieving an acceptably small MBE using this non mass-conserving procedure would require extremely small time steps and would result in excessive calculation time. The MBE for the two mass-conserving procedures P1

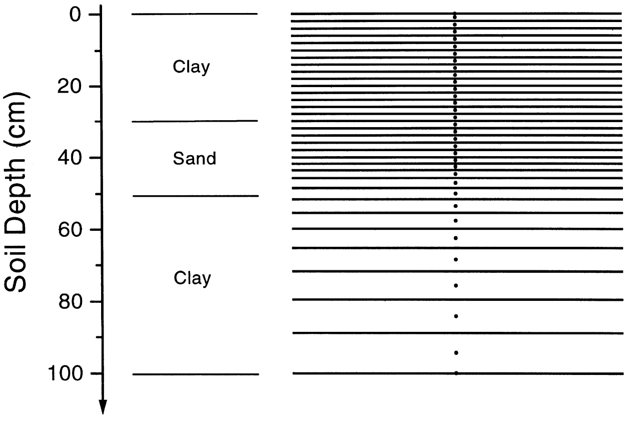

Fig. 4. Inhomogeneous deviation into control volumes of the soil column with three horizons used in Tests 3 and 4.

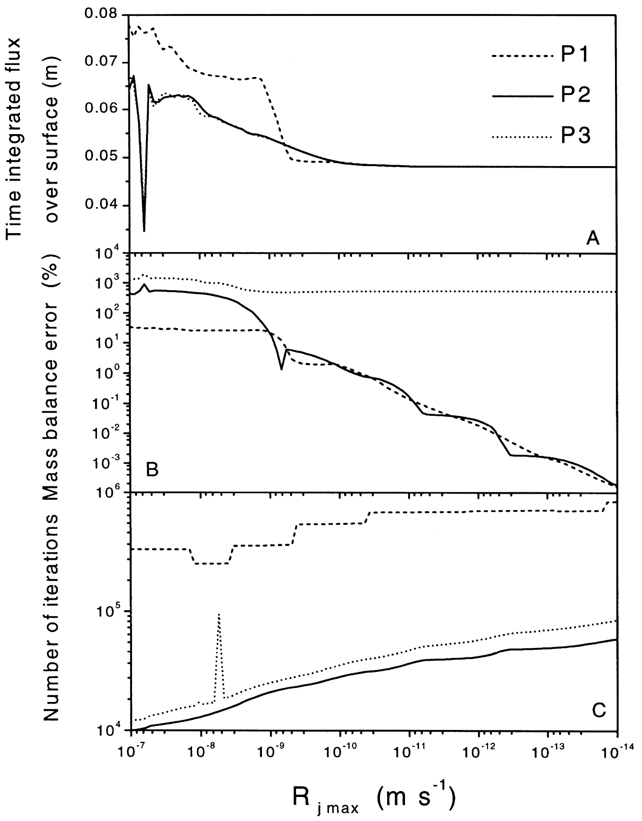

Fig. 5. Test 3: 1 year simulation of the layered soil where the upper boundary condition is given as a highly dynamic flux. Three characteristic parameters are shown as a function of the stop criterion in the iteration processes.

and P2 was of the same order of magnitude for relevant values of $R_{j\ \mathrm{max}}$ and continued to decrease as expected for decreasing values of $R_{j\ \mathrm{max}}$. The number of iterations used (Fig. 5C) for the two mass-conserving procedures was significantly different. On average, P1 used 19 times more iterations than P2 to simulate this highly dynamic infiltration event. This large difference is partly caused by reductions in time steps in P1 in order to aid convergence, which happened in seven situations through the simulated year when a relatively dry top soil was exposed to precipitation. In three situations the time step was cut back to values between 450 and 1800 s, and in four situations the time step was set back to the minimum allowed time step of 3.6 s and kept there for a short period of time. The initial guess given by the Euler step (eqn (2)) was used as the solution in some of these time steps. The difficulties of P1 in obtaining convergence when steep gradients are present was not seen for P2. It should be noted that no difference between the numerical solutions of P1 and P2 was observed. Comparing the total number of iterations used for P2 and P3 shows that mass conservation can have a stabilizing effect on the iteration process. The full year was simulated in 183 s with P1 and in 10 s with P2 on a Pentium/233 MHz personal computer ($R_{j\ \mathrm{max}} = 10^{-12}$ m s^{-1}).

3.4 Test 4

In this test the ability to handle source terms was compared for the two mass-conserving procedures. The starting point was the same problem as simulated in Test 3, but now simulated with the upper boundary condition specified as a known water potential. This water potential at the top of the soil that was calculated in Test 3 (with P2) was stored with one value per hour of simulated time and used as an input in this test. As shown below, this boundary condition gave very similar performances for P1 and P2 when the source term equaled zero. The function in the SOIL model[9] for water uptake by roots was used as an example of a source term. The function is given by the following equations, and shown on Fig. 6A

$$\log(-100\psi) \leq 2.3: \quad S = S_{\mathrm{max}}18 \tag{18}$$

$$2.3 < \log(-100\psi) \leq 4.1: \quad S = S_{\mathrm{max}}\left(\frac{\psi_c}{\psi}\right)^{\beta}$$

$$4.1 < \log(-100\psi) \leq 4.2:$$

$$S = S_{\mathrm{max}}\left(\frac{\psi_c}{\psi_{cc}}\right)^{\beta}(42 - 10\log(-100\psi))$$

where $\beta = 0.5$, $\log(-100\psi_c) = 2.3$, $\log(-100\psi_{cc}) = 4.1$ and S_{max} is the water uptake in a wet soil.

In addition to the explicitly included source term in P1, an implicitly imposed source term in P1 also was tested. The year was simulated repeatedly with increasing values of S_{max} and a constant value of $R_{j\ \mathrm{max}}$ of 10^{-12} m s^{-1}. The results are shown in Fig. 6B.

Note first that the numbers of iterations used in the two procedures were almost identical for S_{max} equal to zero. For increasing values of S_{max} until a value of approximately 0.006 cm^3 cm^{-3} day^{-1} was reached, the number of iterations used in P1 increased evenly, whether the source term was imposed implicitly or explicitly. For larger values of S_{max},

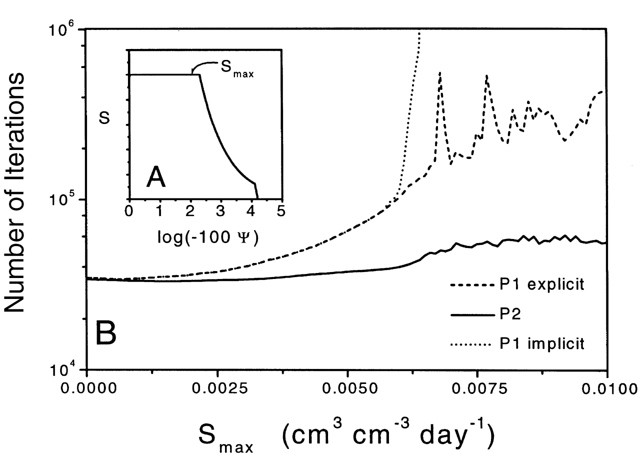

Fig. 6. Test 4: 1 year simulation of the layered soil where a source term is specified. (A) The dependency between the source term and soil water potential (after Jansson[9]). (B) Number of iterations used as a function of the maximum strength of the source term.

the implicitly imposed source term caused divergence (number of iterations $> 2\,000\,000$), and it was only possible to find a solution with P1 if the source term is imposed explicitly. The number of iterations continued to increase for P1, and at $S_{\max}$ equal to 0.01 cm^3 cm^{-3} day^{-1}, the total increase equaled a factor of 13. At several instances the time step was reduced in P1 in order to aid convergence, but only to a minimum value of 1400 s when the source term was imposed explicitly. A much smaller increase equal to a factor of 1.6 was observed for P2 when $S_{\max}$ varied from 0 to 0.01 day^{-1}, compared to the factor of 13 for P1.

4 CONCLUSION

The test results show clearly that the mass-conserving procedures P1 and P2 are superior to the non mass-conserving procedure. Mass conservation can be obtained without any cost in terms of a more complicated implementation, less flexibility, or increased calculation time. In fact, mass conservation seems to have a stabilizing effect on the iteration process (see Fig. 2C and Fig. 5C).

In terms of implementation, P1 is slightly more complicated to compute than P2. This is primarily due to (1) the possible change in time steps in P1 where an ongoing iteration process is terminated and started again with a smaller time step, and (2) the Euler step, which is used as an initial guess in every time step, or in extreme situations, as a solution when a convergent solution cannot be found. As a result of the gradually increasing underrelaxation technique, P2 works more simply with a fixed time step and where no initial guess is calculated. Furthermore, no substituting solution is needed for P2, since the gradually increasing underrelaxation technique will always lead to a convergent solution. This may seem to be a rather arbitrary postulate, but P2 has been tested extensively and divergence has never been observed. It should also be noted that all tests presented in this paper are based on a maximum value of the underrelaxation factor (α^m) of 0.9, which easily can be given a higher value to increase stability.

The observed differences in performance between the two mass-conserving procedures are strongly dependent on the type of boundary condition specified. In all situations where no source terms were involved, and where boundary conditions were given as known water potentials, the number of iterations needed to find a solution was similar for P1 and P2. In more challenging situations, where fluctuating boundary conditions were specified as known fluxes, P1 and P2 performed very differently. In these situations, the time step was occasionally set back in P1, resulting in a large number of iterations used per unit time. The gradually increasing underrelaxation technique allowed P2 to work with the same time step without any noticeable increase in the number of iterations used to find a solution. This resulted in a significant difference (a factor of 300 in Test 2 and a factor of 19 in Test 3) between the iterations needed in the two procedures. The two different approaches for including source terms in P1 and P2 also have a large impact on the performance of the two procedures. The explicitly formulated source term in P1 slowed the rate of convergence for this procedure much more than the approach in P2. This approach also allowed the source term to be included implicitly, which corresponds to the rest of the implicit scheme. The differences in performance between the two mass-conserving procedures P1 and P2 are particularly important in long-term simulations of realistic field situations where calculation time can be crucial.

REFERENCES

1. Ahuja, L. R. & Hebsom, C., Root zone water quality model. GPSR Report 2, USDA, Agriculture Research Service, Fort Collins, CO, 1992.
2. Allen, M. B. and Murphy, C. L. A finite-element collocation method for variably saturated flow in two space dimensions. *Water Resour. Res.*, 1986, **22,** 1537–1542.
3. Celia, M. A., Bouloutaas, E. T. and Zarba, R. L. A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, 1990, **26,** 1483–1496.
4. Celia, M. A., Ahuja, L. R. and Pinder, G. F. Orthogonal collocation and alternating-direction procedures for unsaturated flow problems. *Adv. Water Resour.*, 1987, **10,** 178–187.
5. Cooley, R. L. Some new procedures for numerical solution of variably saturated flow problems. *Water Resour. Res.*, 1983, **19,** 1271–1285.
6. Hansen, S., Jensen, H. E., Nielsen, N. E. and Svendsen, H. Simulation of nitrogen dynamics and biomass production in winter wheat using the Danish simulation model DAISY. *Fert. Res.*, 1991, **27,** 245–259.
7. Haverkamp, R. and Vauclin, M. A note on estimating finite difference interblock hydraulic conductivity values for transient unsaturated flow problems. *Water Resour. Res.*, 1979, **15,** 181–187.
8. Haverkamp, R., Vauclin, M., Touma, J., Wierenga, P. J. and Vachaud, G. A comparison of numerical simulation models for one-dimensional infiltration. *Soil Sci. Soc. Am. J.*, 1977, **41,** 285–294.
9. Jansson, P.-E., Simulation model for soil water and heat conditions. Description of the SOIL model. Tech. Rep. 165, Swedish University of Agricultural Sciences, 1991.
10. Jansson, P.-E. & Halldin, S., Soil water and heat model. Technical description. Swedish Coniferous Forest Project. Tech. Rep. 26, Swedish University of Agricultural Sciences, 1980.
11. Johnsen, K. E., Liu, H. H., Dane, J. H., Ahuja, L. R. and Workman, S. R. Simulating fluctuating water tables and tile drainage with Modified root zone water quality model and a new model WAFLOWM. *Trans. ASAE*, 1995, **38,** 75–83.
12. Johnsson, H., Bergstroem, L., Jansson, P.-E. and Paustian, K. Simulated nitrogen dynamics and losses in a layered agricultural soil. *Agric. Ecosyst. Environ.*, 1987, **18,** 333–356.
13. Milly, P. C. D. Advances in modeling of water in the unsaturated zone. *Transport in Porous Media*, 1988, **3,** 491–514.
14. Milly, P. C. D. A mass-conservative procedure for time-stepping in models of unsaturated flow. *Adv. Water Resour.*, 1985, **8,** 32–36.
15. Milly, P. C. D. & Eagleson, P. S., The coupled transport of water and heat in a vertical soil column under atmospheric excitation. Tech. Rep. 258, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, 1980.

16. Parlange, J.-Y., Barry, D. A., Parlange, M. B., Hogarth, W. L., Haverkamp, R., Ross, P. J., Ling, L. and Steenhuis, T. S. New approximate analytical technique to solve Richards' equation for arbitrary surface boundary conditions. *Water Resour. Res.*, 1997, **33,** 903–906.
17. Patankar, S. V., *Numerical Heat Transfer and Fluid Flow*. McGraw Hill, New York, 1980.
18. Richards, L. A. Capillary conduction of liquids through porous mediums. *Physics*, 1931, **1,** 318–333.
19. Ross, P. J. and Parlange, J.-Y. Comparing exact and numerical solutions of Richards' equation for one-dimensional infiltration and drainage. *Soil Sci.*, 1994, **157,** 341–344.
20. Schnabel, R. R. and Richie, E. B. Calculation of internodal conductances for unsaturated flow simulations: A comparison. *Soil Sci. Soc. Am. J.*, 1984, **48,** 1006–1010.
21. Warrick, A. W. Numerical approximations of darcian flow through unsaturated soil. *Water Resour. Res.*, 1991, **27,** 1215–1222

APPENDIX A

The numerical scheme in procedure P2 is derived below using a straightforward control volume approach.[17] This approach has been used in solving a wide range of extremely complicated problems in the field of modeling fluid flow and heat transfer, and it can also be applied to water movement in soils. One attractive characteristic of the control volume formulation is that it has some intuitive advantages over the finite difference and finite element methods, which makes it easier to evaluate each of the discrete approximations that together form the numerical procedure.

Using the control volume formulation, the soil column is divided into a finite number of non-overlapping control volumes, each containing a grid point at its center. It is assumed that the variation in space and time of the relevant variables is described by piecewise continuous profiles, which are uniquely determined when the grid point values of the variables are known. A general discretization equation is then derived using these profiles in an integration of Richards' equation[18] over a time step and a control volume.

The starting point is Richards' equation where a source term is included

$$\frac{\partial\theta}{\partial t}=\frac{\partial}{\partial z}\left(K\left(\frac{\partial\psi}{\partial z}-1\right)\right)+S \tag{A1}$$

The left side of eqn (A1) can be expressed as

$$\frac{\partial\theta}{\partial t}=\left(\frac{d\theta}{d\psi}\right)\frac{\partial\psi}{\partial t} \tag{A2}$$

where $\mathrm{d}\theta/\mathrm{d}\psi$ is the specific water capacity.

The source term is in many practical applications a nonlinear function of the dependent variable ψ and it is an advantage to directly include at least some of this dependency by expressing S as a linear function of ψ

$$S=S_c+S_p\psi. \tag{A3}$$

The general technique for linearizing the source term (calculate values of S_c and S_p) for a given dependency between S and ψ is outlined below.

Inserting eqns (A2) and (0) in eqn (A3) and integrating the result over the time step from time t to $t+\Delta t$ and over the control volume j from $z_{j-1/2}$ to $z_{j+1/2}$ gives

$$\begin{aligned}&\int_t^{t+\Delta t}\int_{z_{j-1/2}}^{z_{j+1/2}}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)\frac{\partial\psi}{\partial t}\mathrm{d}z\,\mathrm{d}t\\&=\int_t^{t+\Delta t}\int_{z_{j-1/2}}^{z_{j+1/2}}\left(\frac{\partial}{\partial z}\left(K\left(\frac{\partial\psi}{\partial z}-1\right)\right)+S_c+S_p\psi\right)\mathrm{d}z\,\mathrm{d}t\end{aligned} \tag{A4}$$

It is now assumed that $\mathrm{d}\theta/\mathrm{d}\psi$ is constant throughout the time step and that the grid point values of $\mathrm{d}\theta/\mathrm{d}\psi$, $\partial\psi/\partial t$, S_c and $S_p\psi$ prevail throughout the control volume as representative mean values. These assumptions lead to

$$\begin{aligned}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)_j(\psi_j^{n+1}-\psi_j^n)\Delta z_j=&\int_t^{t+\Delta t}\left(\left(K\left(\frac{\partial\psi}{\partial z}-1\right)\right)_{j+1/2}\right.\\&\left.-\left(K\left(\frac{\partial\psi}{\partial z}-1\right)\right)_{j-1/2}+(S_{cj}+S_{pj}\psi_j)\Delta z_j\right)\mathrm{d}t\end{aligned} \tag{A5}$$

where ψ_j^n is the old (known) grid point value of ψ at time t, and ψ_j^{n+1} is the new (unknown) grid point value of ψ at time $t+\Delta t$.

The two first terms in the time integral in eqn (A5) express the Darcy fluxes out and in of the control volume. Assuming that K_{j-1} prevails throughout control volume $j-1$, that K_j prevails throughout control volume j, and that quasi-stationary conditions are present, the flux $q_{j-1/2}$ over the surface of $j-1/2$ can be expressed exactly as

$$\begin{aligned}q_{j-1/2}&=-K_{j-1}\left(\frac{\psi_{j-1/2}-\psi_{j-1}}{0.5\Delta z_{j-1}}-1\right)\\&=-K_j\left(\frac{\psi_j-\psi_{j-1/2}}{0.5\Delta z_j}-1\right)\end{aligned} \tag{A6}$$

where $\psi_{j-1/2}$ is the soil water potential at the surface separating the two control volumes. Eliminating $\psi_{j-1/2}$ from eqn (A6) gives

$$q_{j-1/2}=-K_{j-1/2}\left(\frac{\psi_j-\psi_{j-1}}{0.5(\Delta z_{j-1}+\Delta z_j)}-1\right) \tag{A7}$$

where

$$K_{j-1/2}=\frac{(\Delta z_{j-1}+\Delta z_j)K_{j-1}K_j}{\Delta z_{j-1}K_j+\Delta z_jK_{j-1}} \tag{A8}$$

$K_{j-1/2}$ is the weighted hydraulic conductivity calculated as the well-known harmonic mean, where a non-uniform separation of a soil column into control volumes is taken into account. A large selection of different weighting procedures exist and have been tested and compared in several studies.[7,20,21] In general, the geometric mean ($K_{j-1/2}=(K_{j-1}K_j)^{1/2}$) is considered a better choice than the harmonic mean,[7,20] but it should be noted that the geometric mean cannot be derived from physical arguments as the harmonic mean above.

Inserting eqn (A7) and the equivalent expression for $q_{j+1/2}$ in eqn (A5) gives

$$\begin{aligned}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)_j &(\psi_j^{n+1} - \psi_j^n)\Delta z_j \\ &= \int_t^{t+\Delta t}\Bigg(K_{j+1/2}\left(\frac{\psi_{j+1}-\psi_j}{0.5(\Delta z_j+\Delta z_{j+1})}-1\right) \\ &\quad - K_{j-1/2}\left(\frac{\psi_j-\psi_{j-1}}{0.5(\Delta z_{j-1}+\Delta z_j)}-1\right) \\ &\quad + (S_{cj}+S_{pj}\psi_j)\Delta z_j\Bigg)\,\mathrm{d}t \end{aligned} \tag{A9}$$

In the approximation of the time integral, it is necessary to assume how all the time-dependent terms are varying from time t to $t+\Delta t$. In general, the time integral of $K_{j+1/2}\psi_{j+1}$, for example, can be expressed as

$$\begin{aligned}&\int_t^{t+\Delta t} K_{j+1/2}\psi_{j+1}\,\mathrm{d}t \\ &= ((1-\beta)K^n_{j+1/2}\psi^n_{j+1} + \beta K^{n+1}_{j+1/2}\psi^{n+1}_{j+1})\Delta t\end{aligned} \tag{A10}$$

where β is a weighting factor. Different known schemes can be obtained depending on the value of β. Setting $\beta = 0$ gives an explicit scheme, $\beta = 1/2$ leads to the well-known Cranck–Nicolson scheme, and $\beta = 1$ gives an implicit scheme. Here, the implicit scheme is chosen by which eqn (A9) yields

$$\begin{aligned}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)_j (\psi_j^{n+1}-\psi_j^n)\Delta z_j &= \Delta t K^{n+1}_{j+1/2}\left(\frac{\psi^{n+1}_{j+1}-\psi^{n+1}_j}{0.5(\Delta z_j+\Delta z_{j+1})}-1\right) \\ &\quad - \Delta t K^{n+1}_{j-1/2}\left(\frac{\psi^{n+1}_j-\psi^{n+1}_{j-1}}{0.5(\Delta z_{j-1}+\Delta z_j)}-1\right) \\ &\quad + (S^{n+1}_{cj}+S^{n+1}_{pj}\psi^{n+1}_j)\Delta t\Delta z_j\end{aligned} \tag{A11}$$

The essential element in achieving mass conservation is the approximation of $(\mathrm{d}\theta/\mathrm{d}\psi)_j$. In order to guarantee mass conservation in the numerical solution, it is necessary that the principle of conservation is expressed exactly in the discrete formulation as it is in the differential formulation. The control volume approach leads to a simple balance (a change in water content equals a difference in fluxes) expressing the principle of conservation. Consider a time step where a certain change in water content is calculated for a control volume. In a 'reverse' time step, imposing the opposite directed fluxes in and out of the control volume should lead to the same numeric change in water content. This will not be the case using the standard implicit finite difference approach, where $(\mathrm{d}\theta/\mathrm{d}\psi)_j$ is evaluated at time $t+\Delta t$.

The appropriate approximation of $(\mathrm{d}\theta/\mathrm{d}\theta)_j$ can be derived by integrating eqn (A2) as before over the time step from time t to $t+\Delta t$, and over the control volume j, from $z_{j-1/2}$ to $z_{j+1/2}$

$$\int_{z_{j-1/2}}^{z_{j+1/2}}\int_t^{t+\Delta t}\frac{\partial\theta}{\partial t}\,\mathrm{d}t\,\mathrm{d}z = \int_{z_{j-1/2}}^{z_{j+1/2}}\int_t^{t+\Delta t}\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\frac{\partial\psi}{\partial t}\,\mathrm{d}t\,\mathrm{d}z \tag{A12}$$

In the integration of eqn (A4) it was assumed that $\mathrm{d}\theta/\mathrm{d}\psi$ is constant throughout the time step and that the grid point value of $\mathrm{d}\theta/\mathrm{d}\psi$ prevails throughout the control volume as representative mean values. With these assumptions, eqn (A12) gives

$$\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)_j = \frac{\theta_j^{n+1}-\theta_j^n}{\psi_j^{n+1}-\psi_j^n} \tag{A13}$$

This result is equivalent to the approximation derived by Cooley[15] and Milly[14] using a finite element approach.

Letting j vary from 1 to N (the soil column is divided into N control volumes), eqn (A11) together with eqn (A13) describes a tridiagonal system of N non-linear equations with $N+2$ unknown values of ψ_j^{n+1}. The system of equations will be closed by two additional equations expressing the boundary conditions (described below). The non-linearity of the equations is overcome through iterations, where values of θ_j^{n+1}, K_j^{n+1}, S_{cj}^{n+1} and S_{pj}^{n+1} are evaluated from values of ψ_j^{n+1} from the previous iteration step. Using the symbol m for the iteration level, eqn (A11) can be written as

$$A_j^m\psi_{j-1}^{n+1,m+1} + B_j^m\psi_j^{n+1,m+1} + C_j^m\psi_{j+1}^{n+1,m+1} = D_j^m \tag{A14}$$

where

$$\begin{aligned}A_j^m &= \frac{K^{n+1,m}_{j-1/2}}{0.5(\Delta z_{j-1}+\Delta z_j)} \\ C_j^m &= \frac{K^{n+1,m}_{j+1/2}}{0.5(\Delta z_j+\Delta z_{j+1})} \\ B_j^m &= -A_j^m - C_j^m - \left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)_j^m\frac{\Delta z_j}{\Delta t} + S_{pj}^{n+1,m}\Delta z_j \\ D_j^m &= -K^{n+1,m}_{j-1/2} + K^{n+1,m}_{j+1/2} - \left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)_j^m\frac{\Delta z_j}{\Delta t}\psi_j^n - S_{cj}^{n+1,m}\Delta z_j\end{aligned} \tag{A15}$$

The boundary conditions are introduced by defining two additional and infinitely small control volumes, one on top of the soil surface and one directly below the soil column. These control volumes will be numbered 0 and $N+1$. A Dirichlet condition (a known soil water potential) is assigned as an upper boundary condition by giving the following values to the coefficients in eqn (A14) written for control volume number 0

$$B_0^m = 1,\ \ C_0^m = 0,\ \ D_0^m = \psi_t \tag{A16}$$

where ψ_t is the known soil water potential at the upper boundary. As a lower boundary condition the Dirichlet condition is likewise given by

$$A^m_{N+1} = 0,\ \ B^m_{N+1} = 1,\ \ D^m_{N+1} = \psi_b \qquad \text{(A17)}$$

A Neumann condition (known flux) is expressed as an upper boundary condition by letting j equal one in eqn (A7). Recalling that Δz_0 is infinitely small and substituting $q_{1/2}$ by q_t (the known flux at the boundary) leads to

$$B^m_0 = \frac{K^{n+1,m}_{1/2}}{0.5\Delta z_1},\ C^m_0 = -B^m_0,\ \ D^m_0 = q_t - K^{n+1,m}_{1/2} \qquad \text{(A18)}$$

As a lower boundary condition, the Neumann condition is likewise given by

$$A^m_{N+1} = \frac{K^{n+1,m}_{N+1/2}}{0.5\Delta z_N},\ \ B^m_{N+1} = -A^m_{N+1},\ \ D^m_{N+1} = q_b - K^{n+1,m}_{N+1/2} \qquad \text{(A19)}$$

A source is included through $S^{n+1,m}_{cj}$ and $S^{n+1,m}_{pj}$ in eqn (A15). If the source is independent of ψ, it is sufficient only to initiate $S^{n+1,m}_{cj}$, and to do this before the first iteration step. A source dependent of ψ can also be included only through $S^{n+1,m}_{cj}$ but since the dependency of ψ is not directly included, this strategy will in general slow down the convergence. A better strategy is to include the dependency of ψ through $S^{n+1,m}_{pj}$ which will result in a considerably faster numerical solution. In general, a source S_j can be linearized by

$$S_j = S^{n+1,m}_j + \left(\frac{\mathrm{d}S}{\mathrm{d}\psi}\right)^{n+1,m}_j (\psi^{n+1,m+1}_j - \psi^{n+1,m}_j) \qquad \text{(A20)}$$

Relating eqns (A20) and (A3) gives the following general expressions for $S^{n+1,m}_{cj}$ and $S^{n+1,m}_{pj}$

$$S^{n+1,m}_{pj} = \left(\frac{\mathrm{d}S}{\mathrm{d}\psi}\right)^{n+1,m}_j$$

$$S^{n+1,m}_{cj} = S^{n+1,m}_j - \left(\frac{\mathrm{d}S}{\mathrm{d}\psi}\right)^{n+1,m}_j \psi^{n+1,m}_j \qquad \text{(A21)}$$

It is essential that no values of B^m_j equal zero in the solution of eqn (A14). To prevent this, it is necessary to require that all values of $S^{n+1,m}_{pj}$ are less than or equal to zero (as the first three terms in the expression for B^m_j in eqn (A15)). This requirement is not particularly restrictive, since the relevant processes included through the source term generally have negative values of $S^{n+1,m}_{pj}$ when they are linearized.

One additional advantage of mass-conserving procedures is that the MBE that is present during the process of finding a convergent solution in any arbitrary iteration step can be calculated exactly and used as a criterion for convergence. Recall that eqn (A14) is a balance between the change in water storage per unit time and the water fluxes in and out of the control volume. Assuming that the coefficients $A^m_j + 1$, B^{m+1}_j, C^{m+1}_j and D^{m+1}_j are recalculated right after values of $\psi^{n+1,m+1}_j$ are found by solution of eqn (A14), the MBE per unit time for control volume j is equal to

$$R^{m+1}_j = |D^{m+1}_j - A^{m+1}_j \psi^{n+1,m+1}_{j-1} - B^{m+1}_j \psi^{n+1,m+1}_j - C^{m+1}_j \psi^{n+1,m+1}_{j+1}| \qquad \text{(A22)}$$

The maximum value of R^{m+1}_j is used as a criterion for convergence.

In most situations a converged solution of eqn (A14) is reached in a few iterations. However, sometimes oscillations occur that increase the number of iterations needed to find a converged solution, or even cause divergence in the iteration process. This can effectively be prevented if an underrelaxation of the values of $\psi^{n+1,m+1}_j$ is done immediately after these values are calculated by solution of eqn (A14). The following expression is used

$$\psi^{n+1,m+1,u}_j = (1 - \alpha^m)\psi^{n+1,m+1}_j + \alpha^m \psi^{n+1,m}_j \qquad \text{(A23)}$$

where α^m is the relaxation factor with a value between 0 and 1. Immediately after the underrelaxation is done, $\psi^{n+1,m+1}_j$ is given the values of $\psi^{n+1,m+1,u}_j$. It is a good strategy in each time step to first start the underrelaxation after a few iterations because a converged solution most often is found by then. It is also an advantage to start the underrelaxation with a smaller value of α^m and then increase α^m with the number of iterations in the actual time step.

The denominator in eqn (A13) can be close to or equal to zero. This will be the situation, for example, if a steady-state solution is found. In order to achieve a realistic numerical representation of $(\mathrm{d}\theta/\mathrm{d}\psi)^m_j$ under all circumstances, $(\mathrm{d}\theta/\mathrm{d}\psi)^m_j$ can be calculated as follows. Assuming that θ is known as the unique relation F of ψ, either as a continuous function or as a table where interpolations can be made, $(\mathrm{d}\theta/\mathrm{d}\psi)^m_j$ can then be calculated effectively by

$$\left(\frac{\mathrm{d}\theta}{\mathrm{d}\psi}\right)^m_j = \frac{F(\psi^n_j + \Delta\psi_j) - \theta^n_j}{\Delta\psi_j} \qquad \text{(A24)}$$

where

$$\Delta\psi_j = \mathrm{SIGN}(\psi^{n+1,m}_j - \psi^n_j)\ \mathrm{MAX}(|\psi^{n+1,m}_j - \psi^n_j|, \Delta\psi_{\min}) \qquad \text{(A25)}$$

where the function SIGN equals 1 or -1, depending on the sign of the argument, the function MAX equals the maximum value of the two arguments, and $\Delta\psi_{\min}$ is the predefined minimum numeric values of the denominator (SIGN and MAX are standard functions in most programming languages). Note that if $|\psi^{n+1,m}_j - \psi^n_j|$ is greater than $\Delta\psi_{\min}$, eqn (A24) is identical to eqn (A13).

The entire solution procedure in any arbitrary time step is summarized below:

1. Calculate A^m_j, B^m_j, C^m_j and D^m_j ($m = 0$) using eqns (A15), (A16), (A17), (A18) and (A19) and values for $K^{n+1,m}_{j-1/2}$, $(\mathrm{d}\theta/\mathrm{d}\psi)^m_j$, $S^{n+1,m}_{cj}$ and $S^{n+1,m}_{pj}$ from the last iteration in the previous time step.

2. Find $\psi_j^{n+1,m+1}$ by solution of eqn (A14). Underrelax the values of $\psi_j^{n+1,m+1}$ using eqn (A23) if m is greater than a predefined number.
3. Calculate values of $K_j^{n+1,m+1}$ and $\theta_j^{n+1,m+1}$ based on $\psi_j^{n+1,m+1}$. Calculate values of $K_{j-1/2}^{n+1,m+1}$ using the geometric mean, and values of $(\mathrm{d}\theta/\mathrm{d}\psi)_j^{m+1}$ using eqn (A24). Calculate values of $S_{cj}^{n+1,m+1}$ and $S_{pj}^{n+1,m+1}$ based on $\psi_j^{n+1,m+1}$ using eqn (A21).
4. Calculate A_j^{m+1}, B_j^{m+1}, C_j^{m+1} and D_j^{m+1} using eqns (A15), (A16), (A17), (A18) and (A19) and the maximum value of R_j^{m+1} using eqn (A22). If the maximum value of R_j^{m+1} is greater than a predefined criterion of convergence, repeat steps (2), (3) and (4).