Published in Advances in Water Resources, 23, issue 3, 271-301, 1999
which should be used for any reference to this work


# On the primary variable switching technique for simulating unsaturated–saturated flows

H.-J.G. Diersch [a,*], P. Perrochet [b]

[a] *WASY Institute for Water Resources Planning and Systems Research, Waltersdorfer Str. 105, D-12526 Berlin, Germany*
[b] *Centre d'Hydrogéologie, Université de Neuchâtel, Neuchâtel, Switzerland*

**Abstract**

Primary variable switching appears as a promising numerical technique for variably saturated flows. While the standard pressure-based form of the Richards equation can suffer from poor mass balance accuracy, the mixed form with its improved conservative properties can possess convergence difficulties for dry initial conditions. On the other hand, variable switching can overcome most of the stated numerical problems. The paper deals with variable switching for finite elements in two and three dimensions. The technique is incorporated in both an adaptive error-controlled predictor–corrector one-step Newton (PCOSN) iteration strategy and a target-based full Newton (TBFN) iteration scheme. Both schemes provide different behaviors with respect to accuracy and solution effort. Additionally, a simplified upstream weighting technique is used. Compared with conventional approaches the primary variable switching technique represents a fast and robust strategy for unsaturated problems with dry initial conditions. The impact of the primary variable switching technique is studied over a wide range of mostly 2D and partly difficult-to-solve problems (infiltration, drainage, perched water table, capillary barrier), where comparable results are available. It is shown that the TBFN iteration is an effective but error-prone procedure. TBFN sacrifices temporal accuracy in favor of accelerated convergence if aggressive time step sizes are chosen.

*Keywords:* unsaturated–saturated flow; primary variable switching; newton technique; finite elements; time stepping control; benchmarking; capillary barrier

## 1. Introduction

In the modeling of unsaturated–saturated flow processes several alternatives exist for numerically solving the governing balance equations with their non-linear constitutive relationships. The Darcy equation of fluid motion and the fluid mass conservation equation form the physical basis [2]. In the context of unsaturated flow the basic formulation involves both the fluid pressure head $\psi$ and the saturation $s$ as unknown variables. For these two unknowns only one balance equation, the basic Richards Equation [19], is available. To close the mathematical model one constitutive relationship in form of the capillary pressure head-saturation function is additionally needed to convert one variable to the other (and vice versa). Consequently, the modeller has to decide between primary and secondary variables. Depending on such a choice, different modeling approaches result which are mathematically equivalent in the continuous formulation, but their discrete analogs are different.

As a result, three forms of the unsaturated flow equation can be derived: (1) the pressure based ($\psi$)-form, where the primary variable is the pressure head (or the hydraulic head), (2) the saturation-based ($s$)-form, where the saturation (or the moisture content $\theta$) is chosen as the primary variable, and (3) the mixed ($\psi - s$)-form, where both variables are employed and, in solving the discrete equation system, the pressure head is actually used as the primary variable.

Each of the three different forms has its own advantages and drawbacks. The $\psi$-based form can be used for both saturated and unsaturated soils. The pressure head variable is unique and continuous. Models of this type have been extensively used in various applications [15,18,22,23,29,31,32,35,36,38,42,43]. But, it has been shown [1,4,30,44] that the $\psi$-based form can produce significant global mass balance errors unless very small time steps are used. The $\psi$-based approach can be

* Corresponding author. E-mail: h.diersch@wasy.de

improved if the derivation of the moisture capacity term is performed by suited chord slope approximations in replacing analytical derivatives as proposed by Rathfelder and Abriola [37]. However, the numerical differentiation must be prevented if the pressure head difference falls below a specific range and a proper treatment of the derivative term is then required (for instance, resorting to an analytical evaluation). Accordingly, cord slope approximation does not appear as a general and sufficiently robust technique. It shall fail under drastic parameters and initial conditions. Difficulties of this kind were reported by Paniconi and Putti [36].

Some of these difficulties are avoided when using the mixed-form schemes which possess much better properties with respect to accurate mass conservative solutions. Celia et al. [4] solve the mixed form by a modified Picard iteration scheme. Within the iterative procedure the pressure head is used as the primary variable for the solution at a new iteration step. This mixed Picard technique was successfully applied by Simunek et al. [41] Vogel et al. [46] and Ju and Kung [25] for different situations. Fuhrmann [16] and Lehmann and Ackerer [27] enhanced the mixed form by using a Newton iterative scheme instead of the Picard iteration. Lehmann and Ackerer [27] obtained their best results for one-dimensional problems with the mixed form combined with both the modified Picard and the Newton method. Again, the pressure head was chosen as the primary variable.

Numerical schemes based on the *s*-form of the Richards equation are restricted to unsaturated flow conditions because the saturation variable is not unique for saturated regions, where the soil–water diffusivity goes to infinity and a pressure–saturation relationship no longer exists. Note that the saturation is basically a discontinuous variable. On the other hand, Hills et al. [20] have shown that such a saturation-based algorithm can result in significantly improved performances compared to pressure-based methods, especially when applied to very dry heterogeneous soils. To benefit from the good convergence properties of the *s*-form for both saturated and unsaturated conditions Kirkland et al. [26] suggest to use the saturation in the unsaturated zone and the pressure head in the saturated zone. Unfortunately, their approach is not sufficiently general. As noted by Forsyth et al. [13] the scheme introduces complications for heterogeneous systems, is partially explicit in time, and suffers from balance errors at the transition between the saturated and unsaturated zones.

Recently, Forsyth et al. [13] introduced a powerful new idea in the context of saturated–unsaturated flow simulations. It is termed as the primary variable substitution, or primary variable switching technique, and originates from multiphase flow modeling. It effectively handles the appearance and disappearance of phases [34]. In this approach, a full Newton method is used where the different primary variables, namely saturation and pressure, are switched in different regions depending on the prevailing saturation conditions at each node of a mesh. This technique was found to yield rapid convergence in both the unsaturated and saturated zones compared to pressure-based formulations.

In the light of Forsyth et al.'s work [13], primary variable switching appears as a promising technique to speed up the overall solution process and to tackle difficult-to-solve unsaturated–saturated flow problems. The present study follows these ideas. Modifications and improvements of Forsyth et al.'s scheme consist of (1) a powerful predictor–corrector approach with first and second order accuracy, (2) a one-step full Newton approach with only one control parameter to manage the entire solution process in an adaptive time marching scheme, and (3) a rigorous analytical derivation of the Jacobian of the Newton method. In contrast to the predictor–corrector solution control an aggressive target-based time marching scheme, providing an effective but error-prone strategy, is analyzed.

It will be shown that the primary variable switching technique is the most general approach in which mixed forms using either Picard or Newton techniques appear as special cases. The primary variable switching technique is employed for standard 2D and 3D finite elements. However, the matrix assembly procedure is altered for finite elements depending on the occurrence of primary variables. An upstream weighting scheme is introduced for both structured and unstructured meshes of 2D and 3D finite elements. The paper benchmarks these various schemes by means of selected applications to verify the promised efficiency of primary variable switching. Moisture dynamics in homogeneous and layered soils with dry initial conditions, deemed 'tough' infiltration and drainage problems, and capillary barrier simulations under extreme parameter contrasts and very dry initial conditions are studied. Both agreements and discrepancies are found with previous results presented by Celia et al. [4], Van Genuchten [43], Kirkland et al. [26], Forsyth et al. [13], Webb [47], and Forsyth and Kropinski [14]. Further comparative studies for finding the 'best' solution strategy in practical modeling of unsaturated–saturated flows are required.

## 2. Basic equations

The mass conservation equation of a fluid in a variably saturated media [2] is given by

$$S_o \cdot s(\psi)\frac{\partial \psi}{\partial t} + \varepsilon \frac{\partial s(\psi)}{\partial t} + \nabla \cdot \boldsymbol{q} = Q \tag{1}$$

The fluid motion is described by the Darcy equation written in the form

$$\boldsymbol{q} = -K_{\mathrm{r}}(s)\boldsymbol{K}(\nabla h + \chi \boldsymbol{e}) = -K_{\mathrm{r}}(s)\boldsymbol{K}[\nabla \psi + (1+\chi)\boldsymbol{e}] \tag{2}$$

In Eqs. (1) and (2)

| | |
|---|---|
| $h$ | $=\psi + z$, hydraulic (piezometric) head; |
| $\psi$ | pressure head ($\psi > 0$ saturated medium, $\psi \leqslant 0$ unsaturated medium); |
| $s(\psi)$ | saturation ($0 < s \leqslant 1$, $s=1$, if medium is saturated); |
| $\boldsymbol{q}$ | Darcy flux vector; |
| $z$ | elevation above a reference datum; |
| $t$ | time; |
| $S_o$ | $=\varepsilon\gamma + (1-\varepsilon)\Upsilon$, specific storage due to fluid and medium compressibility; |
| $\varepsilon$ | porosity; |
| $\gamma$ | fluid compressibility; |
| $\Upsilon$ | coefficient of skeleton compressibility; |
| $K_\mathrm{r}$ $(s)$ | relative hydraulic conductivity ($0 < K_\mathrm{r} \leqslant 1$, $K_\mathrm{r}=1$, if saturated at $s=1$); |
| $\boldsymbol{K}$ | tensor of hydraulic conductivity for the saturated medium (anisotropy); |
| $\chi$ | buoyancy coefficient including fluid density effects; |
| $\boldsymbol{e}$ | gravitational unit vector; |
| $Q$ | specific mass supply. |

Constitutive relationships are additionally required (1) for the saturation $s$ as a function of the pressure (capillary) head $\psi$, as well as its inverse, the pressure head $\psi$ as a function of the saturation $s$, and (2) for the relative hydraulic conductivity $K_\mathrm{r}$ as a function of either the pressure head $\psi$ or the saturation $s$. The following empirical relationships are used for the present study [2,46].

*Van Genuchten–Mualem parametric model*:

$$s_\mathrm{e} = \begin{cases} \frac{1}{[1+|\alpha\psi|^n]^m} & \text{for } \psi < \psi_\mathrm{a} \\ 1 & \text{for } \psi \geqslant \psi_\mathrm{a} \end{cases} \tag{3}$$

$$K_\mathrm{r} = s_\mathrm{e}^{\frac{1}{2}}\left\{1 - \left[1 - s_\mathrm{e}^{\frac{1}{m}}\right]^m\right\}^2 \tag{4}$$

*Brooks–Corey parametric model*:

$$s_\mathrm{e} = \begin{cases} \frac{1}{|\alpha\psi|^n} & \text{for } \psi < -1/\alpha \\ 1 & \text{for } \psi \geqslant -1/\alpha \end{cases} \tag{5}$$

$$K_\mathrm{r} = s_\mathrm{e}^{\kappa} \tag{6}$$

with the effective saturation

$$s_\mathrm{e} = \frac{s - s_\mathrm{r}}{s_\mathrm{s} - s_\mathrm{r}} \tag{7}$$

in which

| | |
|---|---|
| $s_\mathrm{e}$ | effective saturation; |
| $s_\mathrm{r}$ | residual saturation; |
| $s_\mathrm{s}$ | maximum saturation; |
| $\psi_\mathrm{a}$ | air-entry pressure head, $\psi_\mathrm{a} \leqslant 0$; |
| $\alpha$ | curve-fitting parameter; |
| $n$ | pore size distribution index, $n \geqslant 1$; |
| $m$ | $= 1 - 1/n$, curve fitting parameter (Mualem assumption); |
| $\kappa$ | $= 2/n + l + 2$, curve-fitting parameter; |
| $l$ | pore-connectivity parameter; |

In combining Eqs. (1) and (2) a general mixed form of the Richards equation naturally results, viz.,

$$\begin{aligned} R(s,\psi) = S_o \cdot s(\psi)\frac{\partial\psi}{\partial t} + \varepsilon\frac{\partial s(\psi)}{\partial t} - \nabla \\ \cdot \{K_\mathrm{r}(s)\boldsymbol{K}[\nabla\psi + (1+\chi)\boldsymbol{e}]\} - Q \\ = 0 \end{aligned} \tag{8}$$

which has to be solved either for $\psi$ (and $h$) or $s$. The retention curves (3) or (5) can be used to convert one variable to the other (and vice versa), viz.,

$$\begin{aligned} s &= f(\psi) \\ \psi &= f^{-1}(s) \end{aligned} \tag{9}$$

## 3. Finite element formulation

Let $\Omega \subset \mathbb{R}^D$ and $(0,\ T)$ be the spatial and temporal domain, respectively, where $D$ is the number of space dimension (2 or 3) and $T$ is the final simulation time, and let $\Gamma$ denote the boundary of $\Omega$ , the weak form of the mass balance equation (1) can be written as

$$\begin{aligned} \int_\Omega wS_o s(\psi)\frac{\partial\psi}{\partial t} + \int_\Omega w\varepsilon\frac{\partial s}{\partial t} - \int_\Omega \boldsymbol{q}\cdot\nabla w \\ = \int_\Omega wQ - \int_\Gamma wq_n \end{aligned} \tag{10}$$

and with Eq. (2) as

$$\begin{aligned} \int_\Omega wS_o s(\psi)\frac{\partial\psi}{\partial t} + \int_\Omega w\varepsilon\frac{\partial s}{\partial t} - \int_\Omega \nabla w\cdot[K_\mathrm{r}(s)\boldsymbol{K}\cdot\nabla\psi] \\ = \int_\Omega wQ - \int_\Gamma wq_n - \int_\Omega \nabla w\cdot[K_\mathrm{r}(s)\boldsymbol{K}\cdot(1+\chi)\boldsymbol{e}] \end{aligned} \tag{11}$$

where $w$ is a test function and $q_\mathrm{n}$ corresponds to the normal fluid flux directed positive outward on $\Gamma$.

In the finite element context a spatial semi-discretization $\Omega^h$ of the continuum domain $\Omega$ is achieved by the union of a set of non-overlapping subdomains $\Omega_e$, the finite elements, as

$$\Omega \approx \Omega^h \equiv \bigcup_e \Omega_e \tag{12}$$

On any finite-element domain $\Omega_e$, the unknown variables and dependent coefficients are replaced by a *continuous approximation* that assumes the separability of space and time, thus

$$\begin{aligned} \psi(x_i,t) &\approx \psi^h(x_i,t) = N_I(x_i)\psi_I(t) \\ s(x_i,t) &\approx s^h(x_i,t) = N_I(x_i)s_I(t) \end{aligned} \tag{13}$$

and, respectively,

$$K_\mathrm{r}(x_i,t) \approx K_\mathrm{r}^h(x_i,t) = N_I(x_i)K_{\mathrm{r}I}(t) \tag{14}$$

where $i = 1,\ldots,D$ represents coordinate indices, $I = 1,\ldots,M$ designates nodal indices, $M$ is the total number of nodes, $N_I$ is the nodal basis function, called the trial space, and $x_i$ are the Cartesian spatial coordi-

nates. Note that the summation convention is used for repeated indices. In our study the basis functions $N_I$ are based on $C_0$ (continuous) piece-wise polynomials that are piecewise-continuously differentiable and square integrable (but whose second and higher derivatives need not to exist).

Using the Galerkin-based finite element method where the test function $w$ becomes identical to the trial space $N$, Eq. (11) leads to the following global matrix system of $M$ equations

$$\boldsymbol{O}(\boldsymbol{s})\dot{\boldsymbol{\Psi}} + \boldsymbol{B}\cdot\dot{\boldsymbol{s}} + \boldsymbol{K}(\boldsymbol{s})\cdot\boldsymbol{\Psi} - \boldsymbol{F}(\boldsymbol{s}) = 0 \tag{15}$$

with its components written in indicial notation

$$O_{IJ}(s) = \sum_e \int_{\Omega_e} N_I S_o s(\psi)\delta_{IJ} \tag{16a}$$

$$B_{IJ} = \sum_e \int_{\Omega_e} N_I \varepsilon \delta_{IJ} \tag{16b}$$

$$K_{IJ}(s) = \sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i} K_{\mathrm{r}}(s) K_{ij} \frac{\partial N_J}{\partial x_j} \tag{16c}$$

$$\begin{aligned} F_I(s) \;&= \sum_e \int_{\Omega_e} N_I Q - \sum_e \int_{\Gamma_e} N_I q_n \\ &\quad - \sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i} K_{\mathrm{r}}(s) K_{ij}(1+\chi) e_j \end{aligned} \tag{16d}$$

where the subscripts $I, J = 1, \ldots, M$ denote nodal indices, $i, j = 1, \ldots, D$ are spatial indices of the Cartesian coordinates, and $\delta_{IJ}$ is the Kronecker operator. The superposed dot means differentiation with respect to time $t$. Non-linearities are shown in parentheses. Note that all matrices connected with time derivatives are lumped. This is virtually mandatory for unsaturated problems to ensure smooth and non-oscillatory solutions [4,25]. The system of equations (15) is highly non-linear due to the functional dependence of the constitutive relationships (3)–(6) for the saturation and the relative conductivity.

The discretized form (15) of the Richards equation is based on the mixed formulation (8), where the fluid and medium compressibility $S_o$ relates to the pressure head $\psi$. For unsaturated conditions the compressibility effects are usually neglected. However, we should mention that the explicit introduction of the $S_o$-term leads to a non-conservative form with respect to the fluid and medium compressibility. For unsaturated conditions (at an arbitrary negative pressure) the discretization (15) is unconditionally mass-conservative for a vanishing $S_o$-term only.

## 4. Temporal discretization

For stability reasons only implicit (A-stable) time discretizations are appropriate for the present class of problems. Otherwise, two-step techniques have to be preferred for multidimensional problems. For the present analysis the fully implicit backward Euler (BE) scheme with a first-order accuracy and the semi-implicit non-dissipative trapezoid rule (TR) with a second-order accuracy are enforced.

Denoting the time plane by the superscript $n$, the implicit form of Eq. (15) reads

$$\boldsymbol{O}\left(\boldsymbol{s}^{n+1}\right)\cdot\dot{\boldsymbol{\Psi}}^{n+1} + \boldsymbol{B}\cdot\dot{\boldsymbol{s}}^{n+1} + \boldsymbol{K}\left(\boldsymbol{s}^{n+1}\right)\cdot\boldsymbol{\Psi}^{n+1} - \boldsymbol{F}\left(\boldsymbol{s}^{n+1}\right) = 0 \tag{17}$$

where the time derivatives are approximated, for the BE scheme, by

$$\dot{\boldsymbol{\Psi}}^{n+1} = \frac{\boldsymbol{\Psi}^{n+1} - \boldsymbol{\Psi}^{n}}{\Delta t_n}, \quad \dot{\boldsymbol{s}}^{n+1} = \frac{\boldsymbol{s}^{n+1} - \boldsymbol{s}^{n}}{\Delta t_n} \tag{18}$$

and for the TR scheme, by

$$\begin{aligned} \dot{\boldsymbol{\Psi}}^{n+1} &= \frac{2}{\Delta t_n}\left(\boldsymbol{\Psi}^{n+1} - \boldsymbol{\Psi}^{n}\right) - \dot{\boldsymbol{\Psi}}^{n}, \\ \dot{\boldsymbol{s}}^{n+1} &= \frac{2}{\Delta t_n}\left(\boldsymbol{s}^{n+1} - \boldsymbol{s}^{n}\right) - \dot{\boldsymbol{s}}^{n}. \end{aligned} \tag{19}$$

Inserting Eqs. (18) and (19) into Eq. (17) results in

$$\begin{aligned} \boldsymbol{R}^{n+1}(\boldsymbol{\Psi}, \boldsymbol{s}) = &\left(\frac{\sigma\boldsymbol{O}(\boldsymbol{s}^{n+1})}{\Delta t_n} + \boldsymbol{K}\left(\boldsymbol{s}^{n+1}\right)\right)\cdot\boldsymbol{\Psi}^{n+1} + \frac{\sigma\boldsymbol{B}}{\Delta t_n} \\ &\cdot\boldsymbol{s}^{n+1} \\ &- \boldsymbol{O}\left(\boldsymbol{s}^{n+1}\right)\left(\frac{\sigma}{\Delta t_n}\boldsymbol{\Psi}^{n} + (\sigma-1)\dot{\boldsymbol{\Psi}}^{n}\right) \\ &- \boldsymbol{B}\left(\frac{\sigma}{\Delta t_n}\boldsymbol{s}^{n} + (\sigma-1)\dot{\boldsymbol{s}}^{n}\right) - \boldsymbol{F}\left(\boldsymbol{s}^{n+1}\right) = 0 \end{aligned} \tag{20}$$

where the weighting factor $\sigma \in (1, 2)$ is unity for the BE scheme and 2 for the TR scheme. It represents a variety of unsaturated flow models, including the variable switching technique, in the most general discrete form. As seen in Eq. (20) the second-order TR scheme is readily available with little extra work. It only differs from the first-order BE scheme by the acceleration terms $\dot{\boldsymbol{\Psi}}^{n}$ and $\dot{\boldsymbol{s}}^{n}$ at the previous time plane, and by the factor $2/\Delta t_n$ instead of $1/\Delta t_n$.

## 5. Primary variable switching methodology

To solve the basic matrix system (20) one has to decide which variable of $\psi(h)$ or $s$ should be primary. Commonly, the selection of the primary variable is done in a static manner and results in a 'fixed' $\psi$-, $s$- or $(\psi - s)$-modeling strategy, including the limitations and drawbacks discussed above. In contrast, primary variable switching is done dynamically depending on the current flow characteristics.

Let $X_I$ be the primary variable associated with node $I$. $X_I$ can be either $\psi_I$ or $s_I$. Accordingly, we can consider $\boldsymbol{X}$ as a vector containing the different primary variables in the solution space $\Omega^h$ as

$$\boldsymbol{X} \in (\Psi, \boldsymbol{s}). \tag{21}$$

Hence, the matrix system (20) can be written in the form

$$\boldsymbol{R}^{n+1}(\boldsymbol{X}) = 0 \tag{22}$$

and solved for $X_I$ $(I = 1, \ldots, M)$.

The solution of the non-linear Eq. (22), i.e., the vector of primary variables $\boldsymbol{X}$, is performed by the Newton method, viz.,

$$\boldsymbol{J}^X\left(\Psi_\tau^{n+1},\ \boldsymbol{s}_\tau^{n+1}\right)\Delta\boldsymbol{X}_\tau^{n+1} = -\boldsymbol{R}_\tau^{n+1}(\Psi, \boldsymbol{s}) \tag{23a}$$

with the increment

$$\Delta\boldsymbol{X}_\tau^{n+1} = \boldsymbol{X}_{\tau+1}^{n+1} - \boldsymbol{X}_\tau^{n+1} \tag{23b}$$

and the Jacobian $\boldsymbol{J}^X$ expressed in indicial notation as

$$J_{IJ}^X\left(\Psi_\tau^{n+1},\ \boldsymbol{s}_\tau^{n+1}\right) = \frac{\partial R_I^{n+1}\left(\Psi_\tau^{n+1},\ \boldsymbol{s}_\tau^{n+1}\right)}{\partial X_{\tau J}^{n+1}} \tag{23c}$$

where $\tau$ denotes the iteration number.

The primary variable at any node $I$ is switched for every Newton iteration $\tau$ by using the following method [13]:

IF $\left(s_{\tau I}^{n+1} \geqslant \mathrm{tol}_f\right)$ THEN

Use $\psi_{\tau I}^{n+1}$ as primary variable at node $I$ and solve the Newton statement (23a) as

$$J_{IJ}^{\psi}\left(\Psi_\tau^{n+1},\ \boldsymbol{s}_\tau^{n+1}\right)\Delta\psi_{\tau J}^{n+1} = -R_{\tau I}^{n+1}(\Psi, \boldsymbol{s}) \tag{24}$$

ELSE IF $\left(s_{\tau I}^{n+1} < \mathrm{tol}_b\right)$ THEN

Use $s_{\tau I}^{n+1}$ as primary variable at node $I$ and solve the Newton statement (23a) as

$$J_{IJ}^{s}\left(\Psi_\tau^{n+1},\ \boldsymbol{s}_\tau^{n+1}\right)\Delta s_{\tau J}^{n+1} = -R_{\tau I}^{n+1}(\Psi, \boldsymbol{s}) \tag{25}$$

ELSE

Do not change primary variable for the node $I$ and solve Eq. (24) or Eq. (25) according to the hitherto selected primary variable $\left(\psi_{\tau I}^{n+1}\right.$ or $\left.s_{\tau I}^{n+1}\right)$.

ENDIF

The Newton approach requires continuous derivatives of the Jacobians $\boldsymbol{J}^{\psi}$ and $\boldsymbol{J}^{s}$ with respect to the pressure head $\psi$ and the saturation $s$, respectively. In the present finite element method the variables $\psi$ and $s$ are approximated in a continuous manner according to Eq. (13) if occurring as primary variables and the Jacobians are thus derivable. On the other hand, variable smoothing is necessary if one determines secondary variables from primary variables using the retention curves (3) or (5) under heterogeneous conditions. To do so, element material quantities have to be averaged at nodal patches. In the context of the finite element method, the arithmetic mean appears as a natural smoothing technique and will be preferred here. Such a smoothing technique is analogous to that of deriving continuous Darcy fluxes in heterogeneous porous media as described in Ref. [10].

The switching tolerances $\mathrm{tol}_f$ and $\mathrm{tol}_b$ have to be appropriately chosen. The following requirements are necessary

$$\mathrm{tol}_f < 1, \quad \mathrm{tol}_f \neq \mathrm{tol}_b. \tag{26}$$

The Jacobians $\boldsymbol{J}^X$ can be computed either numerically or analytically. The analytical method is more efficient [27] and will be preferred in the present study. While a perturbation scheme such as the one used by Forsyth et al. [13] requires a pass of $2M$ evaluations, analytical derivatives require only a pass of $M$ evaluations. The elements of the corresponding Jacobians $\boldsymbol{J}^{\psi}\left(\Psi_\tau^{n+1},\ \boldsymbol{s}_\tau^{n+1}\right)$ of Eq. (24) and $\boldsymbol{J}^{s}\left(\boldsymbol{s}_\tau^{n+1},\ \Psi_\tau^{n+1}\right)$ of Eq. (25) are summarized in the Appendices A and B, respectively. Otherwise, the residual $R_{\tau I}^{n+1}(\Psi, \boldsymbol{s})$ at the iterate $\tau$ and node $I$ is independent of the actually used primary variables $X_I$ and is computed according to Eq. (20) in the following way

$$\begin{aligned} -R_{\tau I}^{n+1}(\Psi, \boldsymbol{s}) = &-\left(\frac{\sigma O_{IJ}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\Delta t_n} + K_{IJ}\left(\boldsymbol{s}_\tau^{n+1}\right)\right) \\ &\cdot \psi_{\tau J}^{n+1} - \frac{\sigma B_{IJ}}{\Delta t_n} \cdot s_{\tau J}^{n+1} \\ &+ O_{IJ}\left(\boldsymbol{s}_\tau^{n+1}\right) \cdot \left(\frac{\sigma}{\Delta t_n}\psi_J^n + (\sigma - 1)\dot{\psi}_J^n\right) \\ &+ B_{IJ} \cdot \left(\frac{\sigma}{\Delta t_n} s_J^n (\sigma - 1)\dot{s}_J^n\right) + F_I\left(\boldsymbol{s}_\tau^{n+1}\right). \end{aligned} \tag{27}$$

It has to be noticed here that the variable switching is generally nodewise. This carries consequences in the finite element assembly technique used to construct the Jacobian $\boldsymbol{J}^X$. Traditionally, the assembling process is performed by

$$J_{IJ}^X = \sum_e \int_{\Omega_e} \{\ldots\}_{\forall I, \forall J} \tag{28}$$

in an elementwise fashion where the nodal contributions are added in the global matrix. This can no longer be done if the primary variables appear in a mixed manner in a mesh. If the primary variables are not of the same kind at a current stage, the following nodewise assembly is required

$$J_{IJ}^X = \sum_I \sum_{e \in \eta_I} \int_{\Omega_e} \{\ldots\}_{I, \forall J} \tag{29}$$

where the contributions from an adjacent element patch $\eta_I$ to a node $I$ are added in the global matrix.

The primary variable switching technique can be considered as a most general formulation in which previous solution strategies are encompassed as special cases. Taking the pressure head $\psi$ as primary variable, omitting for simplicity the compressibility term $\boldsymbol{O}(S_o)$ and considering only the fully implicit BE scheme, we obtain from Eq. (24) and Eq. (A1)

$$\begin{aligned} &\left(\boldsymbol{K} + \Psi_\tau^{n+1}\frac{\partial \boldsymbol{K}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\partial \Psi_\tau^{n+1}} + \frac{\boldsymbol{B}}{\Delta t_n}\frac{\partial \boldsymbol{s}_\tau^{n+1}}{\partial \Psi_\tau^{n+1}} - \frac{\partial \boldsymbol{F}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\partial \Psi_\tau^{n+1}}\right) \\ &\left(\Psi_{\tau+1}^{n+1} - \Psi_\tau^{n+1}\right) = -\boldsymbol{K}\Psi_\tau^{n+1} - \frac{\boldsymbol{B}}{\Delta t_n}\left(\boldsymbol{s}_\tau^{n+1} - \boldsymbol{s}^n\right) + \boldsymbol{F}\left(\boldsymbol{s}_\tau^{n+1}\right) \end{aligned} \tag{30}$$

which is the Newton scheme of the mixed ($\psi - s$)-form of the Richards equation [16,27]. Furthermore, the modified Picard scheme for the mixed ($\psi - s$)-form of the Richards equation [4] can be deduced from Eq. (30) by dropping the partial Jacobians of the 2nd and 4th term of the left-hand side of Eq. (30), yielding

$$\left(\boldsymbol{K} + \frac{\boldsymbol{B}}{\Delta t_n}\boldsymbol{C}_\tau^{n+1}\right)\Psi_{\tau+1}^{n+1} = \frac{\boldsymbol{B}}{\Delta t_n}\boldsymbol{C}_\tau^{n+1}\Psi_\tau^{n+1} - \frac{\boldsymbol{B}}{\Delta t_n}\left(\boldsymbol{s}_\tau^{n+1} - \boldsymbol{s}^n\right) + \boldsymbol{F}\left(\boldsymbol{s}_\tau^{n+1}\right) \tag{31}$$

with the moisture capacity (A7) $\boldsymbol{C}_\tau^{n+1} = \partial \boldsymbol{s}_\tau^{n+1}/\partial \Psi_\tau^{n+1}$. Finally, the common $\psi$-based form is easily obtained from Eq. (31) if the saturation terms on the right-hand side are expressed by their derivatives with respect to the pressure head $\boldsymbol{s}_\tau^{n+1} - \boldsymbol{s}^n = \boldsymbol{C}_\tau^{n+1}\left(\Psi_\tau^{n+1} - \Psi^n\right)$

$$\left(\boldsymbol{K} + \frac{\boldsymbol{B}}{\Delta t_n}\boldsymbol{C}_\tau^{n+1}\right)\Psi_{\tau+1}^{n+1} = \frac{\boldsymbol{B}}{\Delta t_n}\boldsymbol{C}_\tau^{n+1}\Psi^n + \boldsymbol{F}\left(\boldsymbol{s}_\tau^{n+1}\right). \tag{32}$$

While the Newton scheme applied to the primary variable switching technique in Eqs. (24) and (25) and to the mixed form (30) is quadratically convergent, the Picard-type solutions (31) and (32) provide only a linearly convergent accuracy. One notes here that the matrix systems of the Newton method (24), (25) and (30) are always unsymmetric, while the Picard schemes in Eqs. (31) and (32) preserve symmetry of the resulting matrix systems.

The derivation of the family of unsaturated flow models presented here clearly differs from the Newton approach put forward by Paniconi et al. [35], Paniconi and Putti [36], and Miller et al. [29] who started from a $\psi$-based approach in a formal mathematical manner. As a result, the second order derivatives of the saturation relationship arising in the computation of the Jacobian appear somewhat questionable from a physical point of view.

## 6. Solution control

### 6.1. Adaptive predictor–corrector one-step Newton (PCOSN) time marching scheme

Generally, the control of the solution of the resulting highly non-linear matrix systems (24) and (25) is a tricky matter. Both the choice of the time step size $\Delta t_n$ and the iteration control of the Newton scheme significantly influence the success and the efficiency of the simulation. Given that the overall solution process should be performed with a minimum of user-specified control parameters, a fully automatic and adaptive time selection strategy is useful for the present class of problems. In this work a predictor–corrector time integrator is used which was originally introduced by Gresho et al. [17] subsequently improved by Bixler [3], and successfully employed for various buoyant groundwater flow problems [5,10]. It monitors the solution process via a local time truncation error estimation in which the time step size is cheaply and automatically varied in accordance with temporal accuracy requirements. It has been proven to be a cost-effective and robust procedure in that the time step size is increased whenever possible and decreased only if necessary.

In the primary variable switching strategy the Newton method plays a central role. The control of the iteration process with a variable time step size can be combined in the following unified procedure. It is well-known that the Newton scheme converges (with a quadratic convergence rate) if (and only if) a good initial guess of the solution is available. In transient situations this is feasible with a proper adaptation of the time step size to the evolving flow characteristics. At a given time stage, a good initial guess of the solution can always be obtained provided the time step is sufficiently small. Now, it can be argued [17] that the required degree of convergence has to be satisfied in just one full Newton iteration per time step. To do so, the time discretization error $\delta$ can also be used as the Newton convergence criterion for the iterate $\tau$. This is called the one-step Newton method where $\delta$ can be seen as an overall error parameter aiming at keeping the time discretization error small.

For the primary variable switching technique the proposed PCOSN time marching scheme consists of the following main working steps.

*STEP 0*: Initialization

Compute the initial acceleration vectors $\dot{\Psi}^0$ and $\dot{\boldsymbol{s}}^0$ from Eq. (17) as

$$[\boldsymbol{O}(\boldsymbol{s}^0) + \boldsymbol{B}\boldsymbol{C}^0] \cdot \dot{\Psi}^0 = -\boldsymbol{K}(\boldsymbol{s}^0) \cdot \Psi^0 + \boldsymbol{F}(\boldsymbol{s}^0) \tag{33}$$

and with

$$\dot{\boldsymbol{s}}^0 = \boldsymbol{C}^0 \cdot \dot{\Psi}^0 \tag{34}$$

where $\boldsymbol{C}^0$ is the initial moisture capacity vector according to Eq. (A7), $\Psi^0$ and $\boldsymbol{s}^0$ are the initial distributions of the pressure head $\psi$ and the saturation $s$, respectively. Furthermore, we choose an initial time step size $\Delta t_0$.

*STEP 1*: Predictor solutions

Explicit schemes of first and second order accuracy in time provide appropriate predictor solutions for the primary variable $\boldsymbol{X}^{n+1}$ (either $\Psi^{n+1}$ or $\boldsymbol{s}^{n+1}$) at the new time plane $n + 1$. We use either the first-order accurate forward Euler (FE) scheme

$$\boldsymbol{X}_p^{n+1} = \boldsymbol{X}^n + \Delta t_n \dot{\boldsymbol{X}}^n \tag{35}$$

or the second-order accurate Adams–Bashforth (AB) scheme

$$\boldsymbol{X}_p^{n+1} = \boldsymbol{X}^n + \frac{\Delta t_n}{2}\left[\left(2 + \frac{\Delta t_n}{\Delta t_{n-1}}\right)\dot{\boldsymbol{X}}^n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\boldsymbol{X}}^{n-1}\right]. \tag{36}$$

Note here that, since $\dot{\boldsymbol{X}}^{n-1}$ is required, the AB formula cannot be applied before the second step ($n = 1$). The prediction has to be started with the FE procedure,

where $\dot{\boldsymbol{X}}^0$ is available from Eqs. (33) and (34). The subscript $p$ indicates the predictor values at the new time plane $n+1$. In the one-step Newton procedure (i.e., $\tau=1$) the resulting non-linear matrix equations Eqs. (24) and (25) are linearized by using the corresponding predictors. Accordingly, the Newton iterates are taken as

$$\boldsymbol{\Psi}_{\tau}^{n+1}=\boldsymbol{\Psi}_{p}^{n+1},\quad \boldsymbol{s}_{\tau}^{n+1}=\boldsymbol{s}_{p}^{n+1} \tag{37}$$

*STEP 2*: Corrector solutions

Depending on the primary variable switching criteria stated above the following matrix systems (24), (25) arise

$$J_{IJ}^{\psi}\left(\boldsymbol{\Psi}_{p}^{n+1},\ \boldsymbol{s}_{p}^{n+1}\right)\Delta\psi_{pJ}^{n+1}=-R_{pI}^{n+1}(\boldsymbol{\Psi},\boldsymbol{s}),$$
$$\Delta\psi_{pJ}^{n+1}=\psi_{J}^{n+1}-\psi_{pJ}^{n+1} \tag{38}$$

to solve the pressure head $\boldsymbol{\Psi}^{n+1}$ or

$$J_{IJ}^{s}\left(\boldsymbol{\Psi}_{p}^{n+1},\ \boldsymbol{s}_{p}^{n+1}\right)\Delta s_{pJ}^{n+1}=-R_{pI}^{n+1}(\boldsymbol{\Psi},\boldsymbol{s}),$$
$$\Delta s_{pJ}^{n+1}=s_{J}^{n+1}-s_{pJ}^{n+1} \tag{39}$$

to solve the saturation $\boldsymbol{s}^{n+1}$, where the (predicted) residual $\boldsymbol{R}_{p}^{n+1}$ in Eqs. (38) and (39) is also evaluated by using the predictor solutions $\boldsymbol{\Psi}_{p}^{n+1}$ and $\boldsymbol{s}_{p}^{n+1}$ applied to the $\tau$-terms in Eq. (27). Note that the predictor of the FE (35) is used for the BE ($\sigma=1$) and that the predictor of the AB (36) is used for the TR ($\sigma=2$) in Eqs. (38) and (39). Accordingly, the predictor–corrector solutions will be called FE/BE and AB/TR scheme, respectively.

*STEP 3*: Updated accelerations

In preparing the data for the next time step the new acceleration vectors $\dot{\boldsymbol{X}}^{n+1}$ are computed for the FE

$$\dot{\boldsymbol{X}}^{n+1}=\frac{1}{\Delta t_n}\left(\boldsymbol{X}^{n+1}-\boldsymbol{X}^{n}\right) \tag{40}$$

by using the BE (18) and for the AB

$$\begin{aligned}\dot{\boldsymbol{X}}^{n+1}&=\frac{2}{\Delta t_n}\left(\boldsymbol{X}^{n+1}-\boldsymbol{X}^{n}\right)-\dot{\boldsymbol{X}}^{n}\dot{\boldsymbol{X}}^{n}\\&=\frac{\Delta t_{n-1}}{\Delta t_n+\Delta t_{n-1}}\left(\frac{\boldsymbol{X}^{n+1}-\boldsymbol{X}^{n}}{\Delta t_n}\right)\\&\quad+\frac{\Delta t_n}{\Delta t_n+\Delta t_{n-1}}\left(\frac{\boldsymbol{X}^{n}-\boldsymbol{X}^{n-1}}{\Delta t_{n-1}}\right)\end{aligned} \tag{41}$$

by modifying the TR (19) according to Bixler [3].

*STEP 4*: Error estimation

The local truncation error of the approximate equations depends on the predicted $\boldsymbol{X}_{p}^{n+1}$ and corrected $\boldsymbol{X}^{n+1}$ solutions. For the FE/BE and the AB/TR the error estimation yields [17]

$$\boldsymbol{d}^{n+1}=\phi\left(\boldsymbol{X}^{n+1}-\boldsymbol{X}_{p}^{n+1}\right) \tag{42a}$$

with

$$\phi=\begin{cases}\frac{1}{2} & \text{for FE/BE}\\ \frac{1}{3(1+(\Delta t_{n-1}/\Delta t_n))} & \text{for AB/TR}\end{cases}. \tag{42b}$$

Appropriate error norms are applied for the vector $\boldsymbol{d}^{n+1}$. Commonly, the weighted RMS $L_2$ error norm

$$\left\|\boldsymbol{d}^{n+1}\right\|_{L_2}=\left[\frac{1}{M}\left(\sum_{I}^{M}\left|\frac{d_I^{n+1}}{X_{\max}^{n+1}}\right|^2\right)\right]^{1/2} \tag{43}$$

and the maximum $L_\infty$ error norm

$$\left\|\boldsymbol{d}^{n+1}\right\|_{L_\infty}=\frac{1}{X_{\max}^{n+1}}\max_{I}\left|d_I^{n+1}\right| \tag{44}$$

are chosen, where $X_{\max}^{n+1}$ is the maximum value of the current primary variable detected at the time plane $n+1$, and used to normalize the solution vector.

*STEP 5*: Tactic of time stepping

The new provisional time step size can be computed by means of the error estimates (42a), (43), (44), the current time step size $\Delta t_{\mathrm{n}}$, and a user-specified error tolerance $\delta$ as [17]

$$\Delta t_{n+1}=\Delta t_n\left(\frac{\delta}{\left\|\boldsymbol{d}^{n+1}\right\|_{L_p}}\right)^{1/\lambda}$$
$$\lambda=\begin{cases}2 & \text{for FE/BE}\\ 3 & \text{for AB/TR}\end{cases}$$
$$p=\begin{cases}2 & \text{for RMS error norm}\\ \infty & \text{for maximum error norm.}\end{cases} \tag{45}$$

The following criteria are used to monitor the progress of the solution:

1. If
$$\Delta t_{n+1}\geqslant\Delta t_n \tag{46a}$$
the current solution $\boldsymbol{X}^{n+1}$ is accurate within the error bound defined by $\delta$ and the increase of the time step is always accepted.
2. If
$$\zeta\Delta t_n\leqslant\Delta t_{n+1}<\Delta t_n \tag{46b}$$
where $\zeta$ is typically 0.85, the solution $\boldsymbol{X}^{n+1}$ is accepted but the time step size is not changed, i.e., $\Delta t_{n+1}=\Delta t_n$.
3. If
$$\Delta t_{n+1}<\zeta\Delta t_n \tag{46c}$$
the solution $\boldsymbol{X}^{n+1}$ cannot be accepted within the required error tolerance $\delta$ and has to be rejected. The proposed new time step size (45) is reduced according to [5]
$$\Delta t_n^{\text{reduced}}=\frac{\Delta t_n^2}{\Delta t_{n+1}}\left(\frac{\delta}{\left\|\boldsymbol{d}^{n+1}\right\|_{L_p}}\right) \tag{46d}$$
and the solution is repeated for the time plane $n+1$ with $\Delta t_n=\Delta t_n^{\text{reduced}}$.

It is important to note that the error tolerance $\delta$ is the only user-specified parameter to control the entire solution process. The starting-up phase is still influenced by the initial time step $\Delta t_0$ which should be kept small. In practice two further constraints for the time step size have shown to be useful. Firstly, the time step should not exceed a maximum measure, i.e., $\Delta t_n\leqslant\Delta t^{\max}$. Secondly, the rate for changing the time step size $\Xi=\Delta t_{n+1}/\Delta t_n$ has also to be limited, i.e., $\Xi\leqslant\Xi_{\max}$ (say 2 or 3). This can help prevent inefficient oscillations in time step size prediction.

The one-step Newton method embedded in the predictor–corrector schemes (FE/BE or AB/ TR) requires the construction and solution of just one linear(ized) system per time step. The unsymmetric linear systems (38) or (39) are solved via a BiCGSTAB iterative solver [45] pre-conditioned by an incomplete Crout decomposition scheme. The preconditioning process automatically provides a suited scaling of the final matrix system. Otherwise, taking the predictor solutions (35) or (36) the derivative terms (A7) and (B7), namely the moisture capacity and inverse moisture capacity terms, respectively, are easily computed by cord slope approximations as summarized in Appendix C.

It should be emphasized that the proposed PCOSN technique controls the overall temporal discretization error via the tolerance $\delta$. At the same time, $\delta$ is enforced as a convergence limit for the Newton method. This error-controlled solution strategy is very different from the target-based time step selection technique which is discussed next.

### 6.2. Target-based full Newton (TBFN) time stepping scheme

Such type of solution strategy is often used in multiphase flow simulation [12,24]. Applying this technique to unsaturated flow problems Forsyth et al. [13] reported a significant increase in performance compared to common (Picard iteration) solution strategies (e.g., up to 10 times faster). In that work the only criterion is the Newton convergence for a possibly large time step size. The step size is determined from a desired change in the variable per time step given by user-specified targets. The target change parameters are often chosen very large to get aggressive time step sizes. The procedure is carried out in the following steps.

*STEP 1*: Perform Newton iteration

With a given time step size $\Delta t_{\mathrm{n}}$ at time plane $n$+1 (at initial time we start with a sufficiently small $\Delta t_0$) we solve for the new Newton iteration $\tau + 1$ either

$$\boldsymbol{J}^{\psi}\left(\Psi_{\tau}^{n+1},\, \boldsymbol{s}_{\tau}^{n+1}\right)\Delta\Psi_{\tau}^{n+1} = -\boldsymbol{R}_{\tau}^{n+1}(\Psi,\boldsymbol{s}),$$
$$\Delta\Psi_{\tau}^{n+1} = \Psi_{\tau+1}^{n+1} - \Psi_{\tau}^{n+1} \tag{47}$$

for the pressure head $\Psi_{\tau+1}^{n+1}$ or

$$\boldsymbol{J}^{\mathrm{s}}\left(\Psi_{\tau}^{n+1},\, \boldsymbol{s}_{\tau}^{n+1}\right)\Delta\boldsymbol{s}_{\tau}^{n+1} = -\boldsymbol{R}_{\tau}^{n+1}(\Psi,\boldsymbol{s}),$$
$$\Delta\boldsymbol{s}_{\tau}^{n+1} = \boldsymbol{s}_{\tau+1}^{n+1} - \boldsymbol{s}_{\tau}^{n+1} \tag{48}$$

for the saturation $\boldsymbol{s}_{\tau+1}^{n+1}$ as primary variable according to the switching criteria stated above. The Newton iterations are repeated until a satisfactory convergence is achieved, such as

$$\left\|\boldsymbol{d}_{\tau}^{n+1}\right\|_{L_p} < \delta \tag{49a}$$

with

$$\boldsymbol{d}_{\tau}^{n+1} = \boldsymbol{X}_{\tau+1}^{n+1} - \boldsymbol{X}_{\tau}^{n+1} \tag{49b}$$

and where $\left\|\boldsymbol{d}_{\tau}^{n+1}\right\|_{L_p}$ can be used as a RMS ($p=2$, Eq. (43)) or maximum ($p=\infty$, Eq. (44)) error norm.

*STEP 2*: Tactic of time stepping at successful Newton convergence

If Newton iterations have converged a new provisional step size $\Delta t_{n+1}$ can be computed in the following way:

$$\Delta t_{n+1} = \Xi \cdot \Delta t_n \tag{50}$$

where $\Xi$ is a time step multiplier. The latter is determined by the minimum ratio of prescribed target change parameters DXWISH (DSWISH for the saturation $\boldsymbol{s}^{n+1}$ and DPWISH for the pressure head $\Psi^{n+1}$) to the Newton correction, namely

$$\Xi = \min_{I}\left[\frac{\mathrm{DXWISH}}{|X_{\tau+1}^{n+1}, I - X_I^n|}\right]. \tag{51}$$

Additionally, it can be useful to constrain both Eq. (50) by a maximum time step size ($\Delta t_{n+1} \leqslant \Delta t_{\max}$) and Eq. (51) by a maximum multiplier ($\Xi \leqslant \Xi_{\max} = 1.1, \ldots, 5$).

*STEP 3*: Tactic of time stepping if Newton iteration fails

The convergence criterion for the Newton method is given by Eq. (49a). If the Newton scheme does not converge within a maximum number of non-linear iterations $\tau \leqslant$ ITMAX (say 12) the current time step has to be rejected. A reduced time step size is then computed by

$$\Delta t_n^{\mathrm{reduced}} = \Delta t_n/\mathrm{TDIV} \tag{52}$$

and the solution process is restarted for the current time plane $n$ + 1, but with $\Delta t_n = \Delta t_n^{\mathrm{reduced}}$. The time step divider TDIV is usually 2 (sometimes a larger value, e.g. 10, can be useful). Additionally, the behavior of the residual $\boldsymbol{R}_{\tau}^{n+1}(\Psi,\boldsymbol{s})$ can be monitored during the iterations. Taking a RMS norm of the residuals at the current $\left\|\boldsymbol{R}_{\tau}^{n+1}\right\|_{L_2}$ and previous stages $\left\|\boldsymbol{R}_{\tau-1}^{n+1}\right\|_{L_2}$ the iterative process is interrupted as soon as the residual stops to decrease $\left\|\boldsymbol{R}_{\tau}^{n+1}\right\|_{L_2} \geqslant \left\|\boldsymbol{R}_{\tau-1}^{n+1}\right\|_{L_2}$ at a certain iterate ($\tau > 1$).

In the TBFN technique the step size is controlled so that the Newton corrections hit, or are less than, the target change parameters DXWISH. It makes use of the fact that the formulation is mass-conservative for an arbitrary implicit time step size. Indeed, this aggressive time stepping control can be very efficient in finding steady-state solutions, if such solutions exist. But in transient situations, it appears as an error-prone strategy in a potential lacking of temporal accuracy, regardless of the good mass-conservative properties of the scheme. In the examples shown below we shall see partly significant differences between the results of the PCOSN and TBFN schemes.

### 6.3. Convergence criterion

An important aspect of the iterative solution via the PCOSN and TBFN schemes is the choice of an appropriate convergence criterion. The one-step Newton approach of the PCOSN assumes a deviatory (change) error measure $\|\boldsymbol{d}^{n+1}\|_{L_p}$ which is a function of $(\boldsymbol{X}^{n+1} - \boldsymbol{X}_p^{n+1})$, *cf.* Eqs. (42a)–(44). The advantage of the PCOSN is that it controls both the truncation and the iteration errors by only one user-specified tolerance $\delta$. To make the TBFN comparable to the PCOSN scheme we use an equivalent deviatory error norm $\|\boldsymbol{d}_\tau^{n+1}\|_{L_p}$ as a function of $(\boldsymbol{X}_{\tau+1}^{n+1} - \boldsymbol{X}_\tau^{n+1})$, *cf.* Eqs. (49a) and (49b). Such a convergence criterion represents a standard test and is commonly used for Newton methods [11].

Other convergence criteria can sometimes be useful. Instead of the deviatory error estimate $\|\boldsymbol{d}_\tau^{n+1}\|_{L_p}$, the residual $\|\boldsymbol{R}_\tau^{n+1}\|_{L_p}$ may be directly controlled. It represents a direct measure of the global mass balance error after terminating the Newton iteration. For instance one can enforce the condition

$$\|\boldsymbol{R}_\tau^{n+1}\|_{L_p} < \delta_2 \|\boldsymbol{F}^{n+1}\|_{L_p} \tag{53}$$

where a second tolerance $\delta_2$ is introduced and an appropriate normalization of the residual (here with respect to the external supply $\boldsymbol{F}^{n+1}$) is required. Such a convergence control would mean that the one-step Newton approach is no more applicable and that the predictor–corrector scheme has to be controlled by both $\delta$ and $\delta_2$, where $\delta$ measures the temporal discretization error and $\delta_2$ measures the global mass balance error. More than one iteration (we need at least two steps) is then required per time step, making the predictor–corrector technique less attractive. Unlike the PCOSN, the TBFN technique has only one control statement (49a) and, of course, it is easy to replace Eq. (49a) by Eq. (53).

In the present study we do not use the condition (53). We shall show that the $\|\boldsymbol{d}_\tau^{n+1}\|_{L_p}$ error norms are sufficient, at least for the examples considered, to ensure the overall evolution of the non-linear process under a small global mass balance error $\|\boldsymbol{R}_\tau^{n+1}\|_{L_2}$. Additionally, we shall observe $\|\boldsymbol{R}_\tau^{n+1}\|_{L_2}$ in our examples and give estimates of the RMS-based integral (total) mass balance error TMBE ($T$) at the final simulation time $T$ in the form

$$\text{TMBE}\ (T) = \frac{\int_{t=0}^{T} \|\boldsymbol{R}_\tau(t)\|_{L_2} \mathrm{d}t}{\int_{t=0}^{T} \|\boldsymbol{F}(t)\|_{L_2} \mathrm{d}t}. \tag{54}$$

Eq. (54) measures the 'accumulated loss' of mass with respect to the total external supply over the entire simulation period (0, $T$). It is an important error measure to assess the results of long-term simulations, e.g., simulations where small residuals are accumulated over long time periods.

## 7. Upstream weighting

Forsyth and Kropinski [14] pointed out the necessity of upstream weighting in unsaturated–saturated problems to avoid spurious local maxima and minima at coarse mesh sizes. Monotonicity considerations were applied to find appropriate evaluation points for the relative conductivity terms depending on the sign of potential differences along discrete spans (element edges). While a central (standard) weighting results from an average of the relative conductivity at the centroids of elements, an upstream weighting is obtained if the evaluation point is shifted upstream in an element. This technique is different from upwind methods commonly used for convection–diffusion equations [7].

Different approaches exist in unsaturated flow modeling for the representation of material properties. Forsyth and Kropinski [14], Simunek et al. [41] or Oldenburg and Pruess [33] prefer a nodal representation, where material interfaces do not coincide with element boundaries and elemental properties have to be averaged. In such an approach upstream weighting points for evaluating the relative conductivity $K_r$ can be directly located between adjacent nodes. Such schemes have proven to be unconditionally monotone [14].

The present upstream weighting method is based on an elemental representation of material properties. We use the following simple procedure to find appropriate upstream weighting points at an element level. In the examples studied below the usefulness and success of this technique will be shown. A theoretical proof of unconditional monotonicity is, however, beyond the scope of this paper.

A central weighting is equivalent to the influence coefficient method using a linear combination of nodal parameters according to Eq. (14), where the nodal basis functions $N_I(x_i) = N_I(\xi, \eta, \zeta)$ are evaluated at the element centroid ($\xi = \eta = \zeta = 0$); $\xi$, $\eta$, and $\zeta$ represent local coordinates of the finite element. Instead of using the central position, we select an upstream position $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ for computing the relative conductivity via Eq. (14). The evaluation point $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ is used for Gauss integration of the matrix terms (16c) and (16d) and is similar to the Gauss-point-based upwind technique proposed by Hughes [21]. To determine the upstream local coordinates $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ in 2D and 3D elements the following method is applied.

Based on the predicted pressure head $\Psi_p^{n+1}$ (or $\Psi_\tau^{n+1}$ for the TBFN scheme) a specific flux can be computed at a central position of an element $e$

$$\boldsymbol{v}_e^{n+1} = -\nabla N_I(0,0,0) \cdot \left[\psi_{pI}^{n+1} + (1+\chi) e_I\right] \tag{55}$$

and, the trajectory of the vector $\boldsymbol{v}_e^{n+1}$ can be easily found. Along the trajectory, in the upstream direction, the

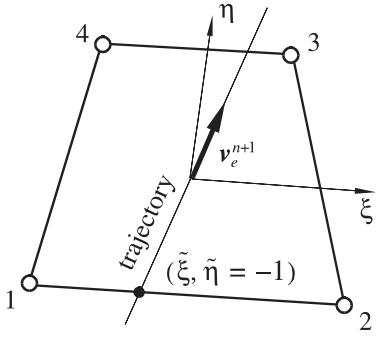


Fig. 1. Upstream local coordinates $(\tilde{\xi}, \tilde{\eta})$ in a 2D finite element.

upstream position $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ is set at the intersection with the element border (Fig. 1).

For the element level $e$ the relative conductivity $K_r(x_i, t) = \bigcup_e K_r^e(x_i, t)$ is evaluated at the upstream point as

$$K_r^e(x_i, t) = N_I(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta}) K_{rI}^e(t) \tag{56}$$

where $K_{rI}^e(t)$ represents the nodal relative conductivities computed as a function of the nodal saturation $s_I(t)$ (or pressure head $\psi_I(t)$). With the upstream point $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ the relative conductivity $K_r^e$ is evaluated only along element edges. For instance, considering the situation in Fig. 1 for a 2D isoparametric finite element, $\tilde{\eta}$ is $-1$ and $K_r^e$, from Eq. (56), becomes independent of nodes 3 and 4, *viz.*, $K_r^e = [(1 - \tilde{\xi}) K_{r1}^e + (1 + \tilde{\xi}) K_{r2}^e]/2$.

## 8. Simulations

The following examples are used to benchmark the primary variable switching technique combined with the PCOSN time marching procedure against traditional and alternative solution strategies. Its efficiency is demonstrated by means of applications where other schemes fail or run eventually into difficulties. The control parameters enforced in these examples are the primary variable switching tolerances (26) [13]

$$\begin{aligned} \mathrm{tol}_f &= 0.99 \\ \mathrm{tol}_b &= 0.89 \end{aligned} \tag{57}$$

and the $\delta$ tolerance encompassing both the time truncation error measure and the Newton convergence criterion is

$$\delta = 10^{-4} \tag{58}$$

using the RMS error norm (43) as the default options. Exceptions will be indicated. Since the proposed schemes are mass-conservative the balance error is a function of the error tolerance $\delta$. This parameter is very important, but its significance with respect to mass balance should not be over-interpreted. As already pointed out by Kirkland et al. [26] a good mass balance does not mean that the distribution of mass across the system has been correctly evaluated. This will be shown in the case of the TBFN time stepping strategy where the following aggressive target change parameters

$$\begin{aligned} \mathrm{DSWISH} &= 0.4 \\ \mathrm{DPWISH} &= 4000\ \mathrm{kPa} \end{aligned} \tag{59}$$

will be used [13]. In the TBFN solution technique temporal non-linear discretization errors may occur due to a fast-but-coarse time stepping. The total mass balance errors will be quantified by the TMBE ($T$) estimate (54).

The large target change parameters (59) were used by Forsyth et al. [13] to illustrate the robustness of the variable switching technique. They did not intend to consider the time truncation errors arising for the large time step sizes generated. Clearly, employing smaller target change parameters would lead to smaller time step sizes and to reduced time truncation errors. But, due to the empirical nature of the control parameters for the TBFN strategy, an optimal parameter choice is not easy and a normal user would likely tend to accept a solution at an 'efficient' time step size as soon as the solution has converged.

It should be noted that spatial discretization errors due to mesh effects are not controlled by $\delta$ (this would require a fully adaptive solution strategy similar to [6] and represents a future challenging problem in unsaturated flow). Instead, spatial discretization effects are analyzed by comparing different mesh resolutions whenever available and appropriate.

### *8.1. Infiltration in homogeneous and inhomogeneous soil columns*

#### *8.1.1. Celia et al.'s problem*

Celia et al. [4] introduced a modified Picard method for the mixed $(\psi - s)$-form of the Richards equation to study water infiltration in a homogeneous soil column with the following parameters [40]: column length of 1 m, Van Genuchten–Mualem parametric model (3), (4) in using $n = 2$, $(m = 0.5)$, $\alpha = 3.35\ \mathrm{m}^{-1}$, $\varepsilon = 0.368$, $s_r = 0.277$, and $s_s = 1.0$, isotropic saturated conductivity of $0.922 \cdot 10^{-4}\ \mathrm{ms}^{-1}$, vanishing compressibility $S_o \approx 0$, zero air-entry pressure head $\psi_a = 0$, constant pressure head $\psi = -0.75$ m at the top and $\psi = -10.0$ m at the bottom, and initial pressure head $\psi^0 = -10.0$ m. We choose an initial time step size of $\Delta t_0 = 10^{-5}$ d. The same spatial discretization characteristics as given in [4] are applied, where $\Delta z = 0.5$ cm (dense grid) and $\Delta z = 2.5$ cm (coarse grid). In [4] dense-grid simulations were performed with a constant time increment of $\Delta t = 60$ s, which means their 'best' solutions for a simulation time of 1 day were obtained after 1440 time steps plus a number of unreported Picard steps.

Fig. 2 compares the pressure profiles computed by the PCOSN scheme with Celia et al.'s solution for the dense grid at a simulation time of 1 day. The agreement

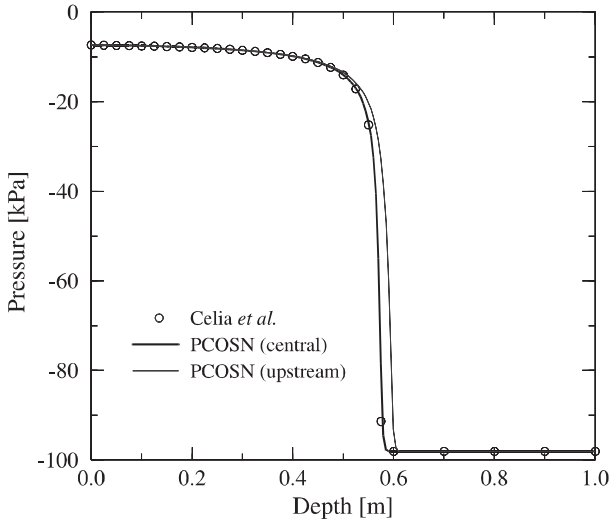


Fig. 2. Pressure profiles at $t = 1$ day for the dense grid: PCOSN results for central and upstream weighting (both FE/BE and AB/TR scheme) with error $\delta = 10^{-4}$ in comparison with Celia et al.'s results [4,40].

is quite perfect if using the standard central weighting scheme. Clearly, for this problem an upstream weighting is numerically not required because the central weighting solutions are non-oscillatory. Nevertheless, if applying upstream weighting a typical phase lead error appears as seen in Fig. 2. It is important to note that the same curves are generated for both the first-order accurate FE/BE and the second-order accurate AB/TR PCOSN schemes. Furthermore, if relaxing the error bound $\delta$ to $10^{-3}$ the FE/BE scheme still gives identical results, but the AB/TR began to fail in producing nonlinear wiggles.

Alternatively, if we use a Newton mixed $(\psi - s)$-form scheme, *cf*. Eq. (30), where the primary variable is always the pressure head $\psi$, with a FE/BE time marching strategy the same results as outlined in Fig. 2 are obtained. However, compared to the PCOSN variable switching, more than thrice the number of Newton steps are required for the same error parameter. Table 1 summarizes the solution effort needed for the different predictor–corrector schemes and error tolerances.

Fig. 3 presents a comparison of the dense and coarse grid solutions to illustrate spatial discretization effects. As shown, a significant phase lead and a somewhat smeared pressure profile result. A similar effect is also obtained if an inappropriate time stepping is selected as displayed in Fig. 4. The TBFN scheme requires only a small number of Newton steps as summarized in Table 2. Solutions were obtained up to five times faster than the PCOSN and up to eighteen times faster than the Newton mixed $(\psi - s)$-form under comparable conditions. The price to pay for that is a remarkable loss of accuracy (Fig. 4). It is important to indicate that this effect is independent of the Newton convergence limit $\delta$. We obtained the same leading curve behavior if decreasing $\delta$ (down to $10^{-6}$). As given in Table 2 the TBFN scheme takes 18 time steps for a constraint of $\Xi_{\max} = 2$. Only when we increase the number of time steps (e.g., enforce an unusual constraint of $\Xi_{\max} = 1.1$) the accuracy improves (*cf*. Fig. 4). This clearly indicates that the error of the TBFN scheme is caused by temporal discretization, which will be further discussed below.

The time behavior of the residual error $\|R\|_{L_2}$ is plotted in Fig. 5 for the TBFN and PCOSN schemes. While the PCOSN terminates with errors in the range of $10^{-5} - 5 \cdot 10^{-7}$, the TBFN produces $\|R\|_{L_2}$ errors smaller than $10^{-6}$ with the limit of $\delta = 10^{-4}$ for a RMS error convergence criterion (43). The total mass balance error TMBE ($T = 1$ d), Eq. (54), can be estimated at $O(10^{-3})$ for the PCOSN and $O(10^{-4})$ for the TBFN.

### *8.1.2. Van Genuchten's problem*

Van Genuchten [43] describes results for moisture movement in a layered soil. A soil column with a length of 170 cm includes 4 layers: clay loam (0–25 cm), loamy sand (25–75 cm), dense material (75–87 cm) and sand

Table 1
Solution effort needed for the PCOSN variable switching scheme compared to the Newton mixed $(\psi - s)$-form solution (dense grid, simulation time 1 day)

| Scheme | Type | Weighting | Error δ | Actual time steps | Total Newton steps [a] | Efficiency |
|---|---|---|---|---|---|---|
| PCOSN | FE/BE | Central | $10^{-4}$ | 437 | 443 | 1.0 |
| PCOSN | FE/BE | Upstream | $10^{-4}$ | 379 | 386 | 0.87 |
| PCOSN | FE/BE | Central | $10^{-3}$ | 283 | 352 | 0.79 |
| PCOSN | FE/BE | Upstream | $10^{-3}$ | 148 | 151 | 0.34 |
| PCOSN | AB/TR | Central | $10^{-4}$ | 436 | 580 | 1.31 |
| PCOSN | AB/TR | Upstream | $10^{-4}$ | 330 | 355 | 0.80 |
| PCOSN | AB/TR | Central | $10^{-3}$ | failed | failed | – |
| PCOSN | AB/TR | Upstream | $10^{-3}$ | failed | failed | – |
| Mixed | FE/BE | Central | $10^{-4}$ | 1406 | 1556 | 3.51 |
| Mixed | FE/BE | Upstream | $10^{-4}$ | 1270 | 1353 | 3.05 |
| Mixed | FE/BE | Central | $10^{-3}$ | 430 | 477 | 1.08 |
| Mixed | FE/BE | Upstream | $10^{-3}$ | 388 | 431 | 0.97 |

[a] Including rejected steps.

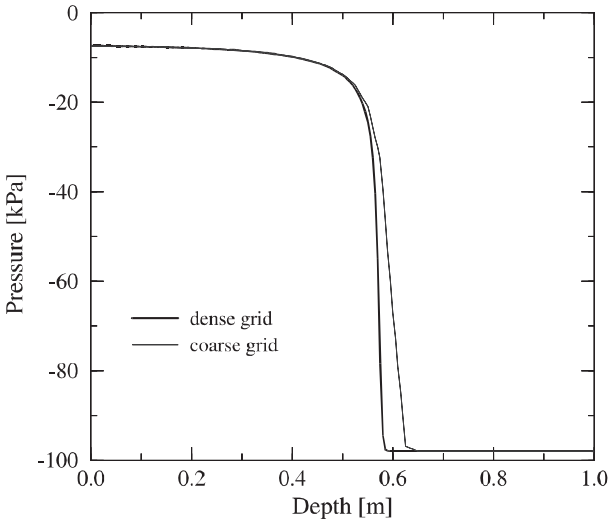


Fig. 3. Pressure profiles at $t = 1$ day computed by the PCOSN scheme (central weighting) with error $\delta = 10^{-4}$ for the dense and coarse grid.

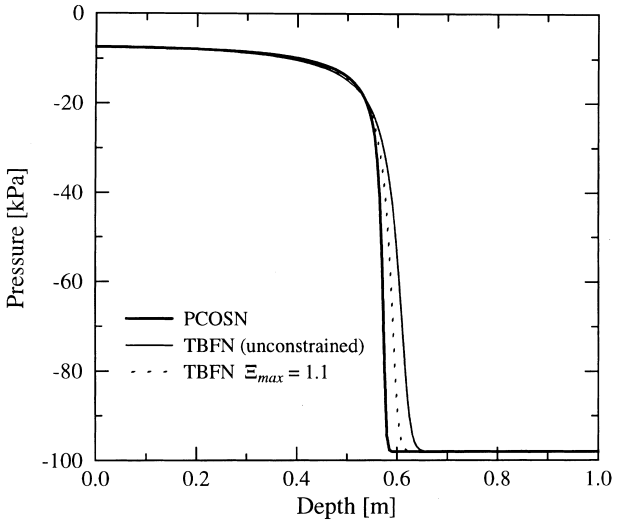


Fig. 4. Computed pressure profiles at $t = 1$ day for the PCOSN scheme (with $\delta = 10^{-4}$) and the TBFN scheme (using $\delta = 10^{-4}, \ldots, 10^{-6}$) at unconstrained ($\Xi_{\max} = \infty$) and constrained ($\Xi_{\max} = 1.1$) time stepping; dense grid and central weighting.

(87–170 cm), where the loamy-sand layer properties change gradually with depth. The initial conditions for the flow are given by $\psi^0 = -3.5$ m. A constant flux is specified at the surface $q_n^h = -0.25$ m/d for $t \leqslant 1$ day (infiltration) and $q_n^h = 0.005$ m/d for $t > 1$ day (evaporation). At the bottom, a drainage gradient-type boundary condition of $q_n^{h\nabla} = K|_{\text{bottom}} = 4$ m/d is imposed [8]. Accordingly, the bottom boundary can freely drain [28]. The parameters of the constitutive relations (Van Genuchten–Mualem model) are fully listed in [40]. The column is discretized in 170 elements, i.e., $\Delta z = 1$ cm. The initial time step is $\Delta t_0 = 10^{-5}$ d.

This problem is not particularly difficult to solve, since the initial conditions are not very dry. All formulations and schemes were successful. Their results are in good agreement with Van Genuchten's solutions as shown in Fig. 6 for the infiltration period. Differences between central and upstream weighting are also exhibited in Fig. 6. To study the merits and solution efforts of the different numerical schemes for this heterogeneous system, let us focus on the saturation profile computed at the end of the infiltration period ($t = 1$ d) under low and extremely high initial suction conditions $\psi^0$.

Using the PCOSN scheme with FE/BE and central weighting the computed saturation profiles at $t = 1$ d is shown in Fig. 7 for different $\psi^0$. As expected, at very dry initial conditions the saturation profile remains unchanged, proving thus the good conservative properties of the variable switching technique. Practically any arbitrary large value of initial suction can be enforced. In contrast to this, standard formulations using the pressure head $\psi$ as primary variable can run into difficulties or completely fail. Especially for very dry conditions there is practically no way to find reasonable convergent solutions in acceptable times. Fig. 8 shows the results for both the mixed ($\psi - s$)-form with Newton iteration (comparable to Eq. (30)) and the standard $\psi$-form with

Table 2
Solution effort for the TBFN scheme using fully implicit time stepping and central weighting (dense grid, simulation time 1 day)

| Error $\delta$ | Constraint $\Xi_{\max}$ | Weighting | Actual time steps | Total Newton steps [a] | Efficiency (Table 1) |
|---|---|---|---|---|---|
| $10^{-4}$ | $\infty$ | Central | 8 | 88 | 0.2 |
| $10^{-4}$ | $\infty$ | Upstream | 5 | 63 | 0.14 |
| $10^{-4}$ | 2 | Central | 18 | 85 | 0.19 |
| $10^{-4}$ | 2 | Upstream | 18 | 94 | 0.21 |
| $10^{-4}$ | 1.1 | Central | 97 | 263 | 0.59 |
| $10^{-4}$ | 1.1 | Upstream | 97 | 309 | 0.70 |
| $10^{-3}$ | 2 | Central | 18 | 65 | 0.15 |
| $10^{-3}$ | 2 | Upstream | 18 | 70 | 0.16 |
| $10^{-5}$ | 2 | Central | 18 | 96 | 0.22 |
| $10^{-5}$ | 2 | Upstream | 18 | 120 | 0.27 |
| $10^{-6}$ | 2 | Central | 18 | 102 | 0.23 |
| $10^{-6}$ | 2 | Upstream | 18 | 143 | 0.32 |

[a] Including rejected steps.

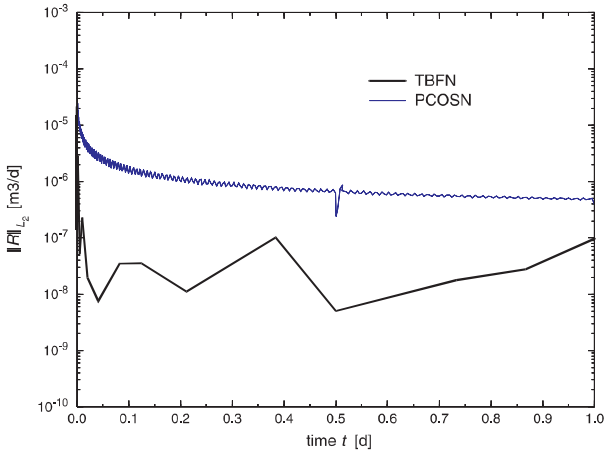


Fig. 5. History of residual error $\|R\|_{L_2}$ for the TBFN and PCOSN schemes with $\delta = 10^{-4}$, RMS error convergence criterion (43) and central weighting.

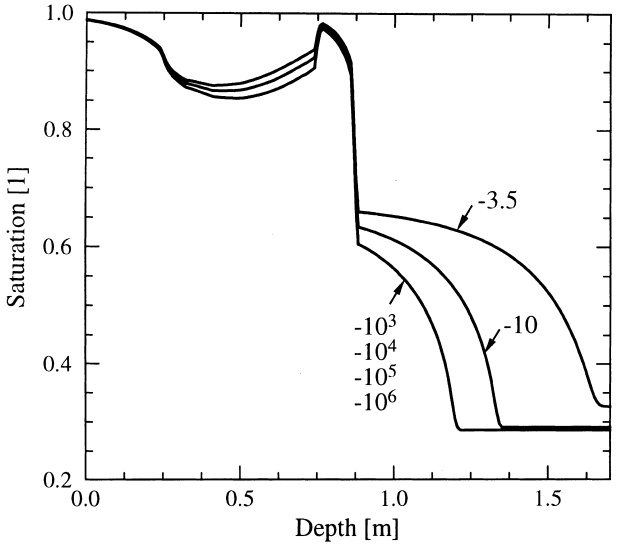


Fig. 7. Saturation distribution at $t = 1$ day computed by the PCOSN scheme (FE/BE, central weighting) with error $\delta = 10^{-4}$ for various initial pressure heads $\psi^0$ in (m).

Picard iteration and cord slope approximation. As seen at low suction ($\psi^0 = -3.5$ m) the schemes yield the same results. However, already for $\psi^0 = -10$ m the standard $\psi$-form reveals mass-conservative problems (phase lag). The phase lag error dramatically grows at higher initial suctions as evidenced in Fig. 8 for $\psi^0 = -10^3$ m. On the other hand, the conservative mixed ($\psi - s$)-form provides better results, though not without a phase lag error at $\psi^0 = -10^3$ m (Fig. 8) in comparison to the good PCOSN results (Fig. 7). We were not able to find convergent solutions for both the mixed ($\psi - s$)-form and the standard $\psi$-form at higher suction values ($\psi^0 < -10^3$ m).

A comparison of the PCOSN and the TBFN variable switching schemes is given in Fig. 9. At low suction values the differences can be seen in the typical lead effects in the saturation profile. This is caused by the poorer temporal accuracy of the TBFN scheme which takes a much smaller number of time steps than the

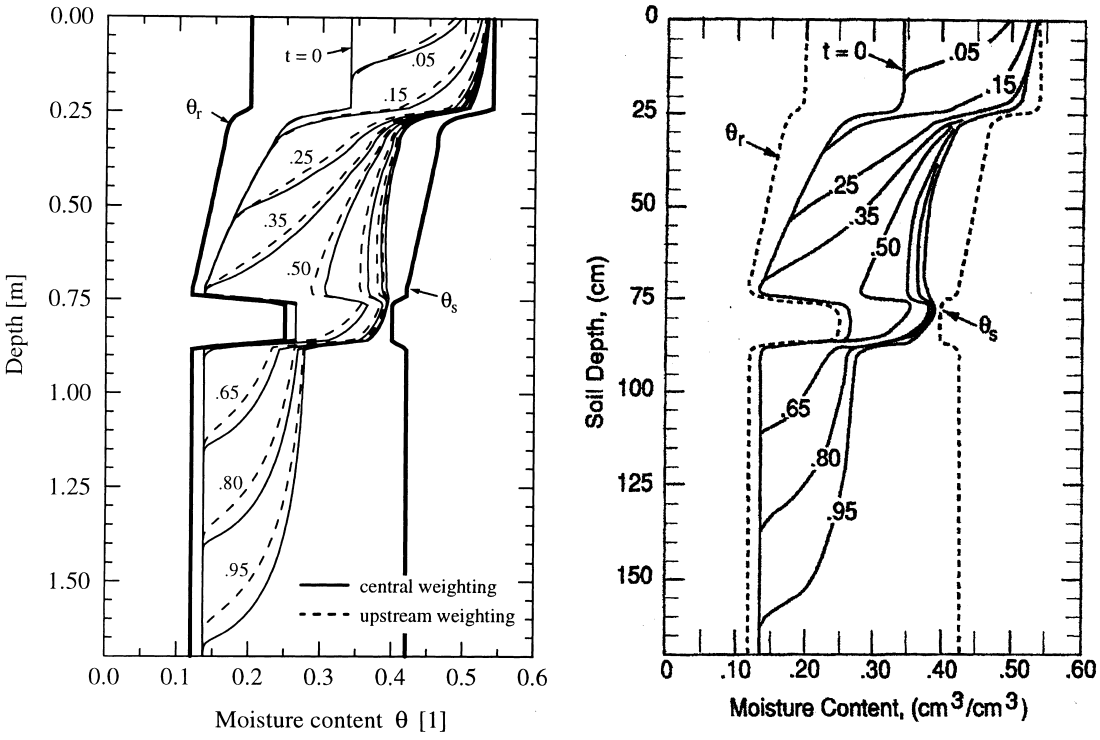


Fig. 6. Simulated moisture-content profiles ($\theta = s \cdot \varepsilon$) during infiltration: present solutions (left) and Van Genuchten's results (right), time in days.

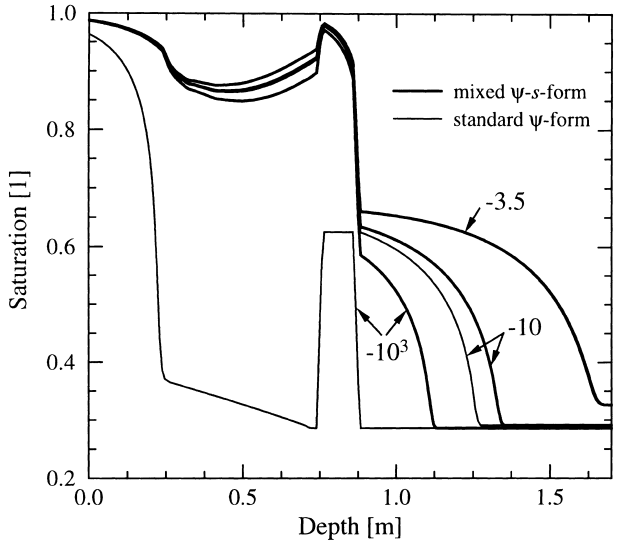


Fig. 8. Saturation distribution at $t = 1$ day computed by the Newton mixed $(\psi - s)$-form and the standard Picard iteration $\psi$-form (FE/BE, central weighting) with error $\delta = 10^{-4}$ for various initial pressure heads $\psi^0$ in (m).

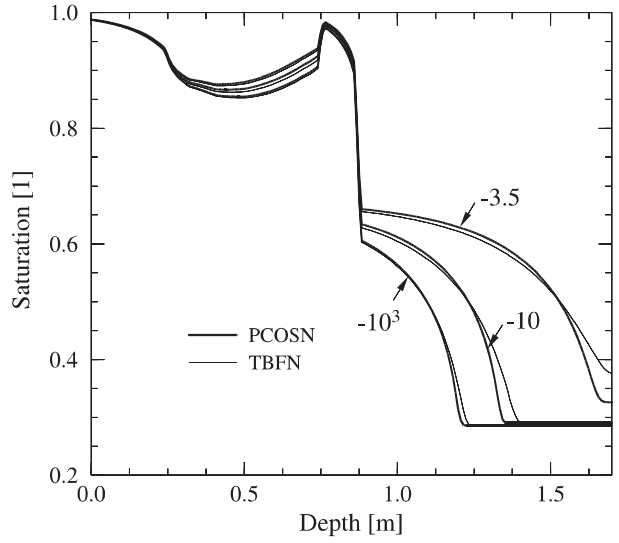


Fig. 9. Comparison of the PCOSN and the TBFN saturation distributions simulated at $t = 1$ day (FE/BE, central weighting) with error $\delta = 10^{-4}$ for various initial pressure heads $\psi^0$ in (m).

error-controlled PCOSN scheme. However, under very dry conditions the differences disappear. For initial pressure heads smaller than $-10^4$ m the computed saturation profiles become identical.

Table 3 summarizes the solution effort in terms of time steps and number of iterations for different schemes depending on the initial suction $\psi^0$. The variable switching techniques (PCOSN and TBFN, columns 2–5 of Table 3 were successful for all $\psi^0$ considered, while the schemes using the pressure head $\psi$ as primary variable (mixed Newton $(\psi - s)$-form with both PCOSN and TBFN, and standard Picard-form, columns 6–11 of Table 3) have shown unsuitable for very dry conditions $\psi^0 < -10^3$ m. The most interesting outcomes of these comparisons are the following.

For variable switching the TBFN scheme is about three to five times faster than the PCOSN scheme. Under very dry conditions the TBFN is definitely superior to PCOSN since the results are virtually equivalent (*cf.* Fig. 9). It should be recalled that the PCOSN scheme is driven by controlling the temporal discretization error while the TBFN scheme is not. The required number of time steps increases naturally with decreasing $\psi^0$. At the same time, the number of rejected steps increases so that the overall effort grows with decreasing $\psi^0$.

The power of the variable switching technique becomes obvious if comparing it with the $\psi$ primary variable solution under the same time stepping strategy. We additionally applied the TBFN technique to the $\psi$ primary variable form, omitting the variable switching. The computational effort dramatically increases by orders of magnitude (3–168 times slower than the TBFN with variable switching as indicated by columns 9 *vs.* 5 of Table 3). Similar observations were made by Forsyth et al. [13]. It is interesting to note that the advantage of the TBFN scheme with respect to the computational effort vanishes for the $\psi$ primary variable form (with the targets (59)). Here, the PCOSN scheme is comparable or even faster (*cf.* columns 7 *vs.* 9 in Table 3). However, the TBFN scheme was able to find convergent solutions for all $\psi^0$, but the required number of Newton steps became extremely large for very dry conditions, unacceptable for practical modeling.

For the variable switching technique we found the following estimates of the total mass balance error TMBE ($T = 1$ d). At lower suction heads $\psi^0$, see Table 3, TMBE ($T = 1$ d) is of O ($10^{-4}$) for the PCOSN and O ($10^{-5}$) for the TBFN. At higher suction heads $\psi^0$ we found TMBE ($T = 1$ d) of O ($10^{-3}$) for the PCOSN and O ($10^{-4}$) for the TBFN.

### *8.2. Drainage of a very coarse material*

The drainage of a very coarse material represents an interesting and challenging test case. By using a $\psi(s)$-curve with no (or negligible) capillarity (very large $\alpha$ in Eq. (3) or Eq. (5)) the medium is at the residual saturation $s_r$ very rapidly and the mass balance can be checked without computing the remaining water in the drained area. The problem is described in Fig. 10. Due to the large $\alpha$-parameters the numerical simulation becomes difficult for an unsaturated–saturated modeling approach (in contrast to a much easier free-surface modeling approach as discussed in Ref. [8]). The problem is solved by using both the Van Genuchten–Mualem

Table 3
Solution effort for different schemes (simulation time 1 day, FE/BE, central weighting, error $\delta = 10^{-4}$, time constraint $\Xi_{max} = 2$)

| Initial pressure head $\psi^0$ (m) | Variable switching | | | | Primary variable $\psi$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PCOSN | | TBFN [a] | | Mixed ($\psi - s$)-form, Newton, Eq. (30) | | | | Standard $\psi$-form, Picard, Eq. (32) | |
| | | | | | PCOSN | | TBFN [a] | | | |
| | Time steps | Total Newton steps [b] | Time steps | Total Newton steps [b] | Time steps | Total Newton steps [b] | Time steps | Total Newton steps [b] | Time steps | Total Picard steps [b] |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| −3.5 | 358 | 360 | 32 | 109 | 634 | 638 | 43 | 292 | 643 | 648 |
| −10 | 676 | 684 | 34 | 171 | 1824 | 2112 | 154 | 1535 | 1760 | 2021 |
| $-10^3$ | 1510 | 2187 | 66 | 580 | 4202 | 4792 | 929 | 9186 | 1128 | 1472 |
| $-10^4$ | 1990 | 3254 | 76 | 673 | Failed | | 1247 | 11535 | Failed | |
| $-10^5$ | 2180 | 3858 | 97 | 831 | Failed | | 1539 | 14138 | Failed | |
| $-10^6$ | 2696 | 4988 | 115 | 952 | Failed | | 155025 | 159641 | Failed | |

[a] Additional time constraint $\Delta t_{max} = 0.05$ d.
[b] Including rejected steps.

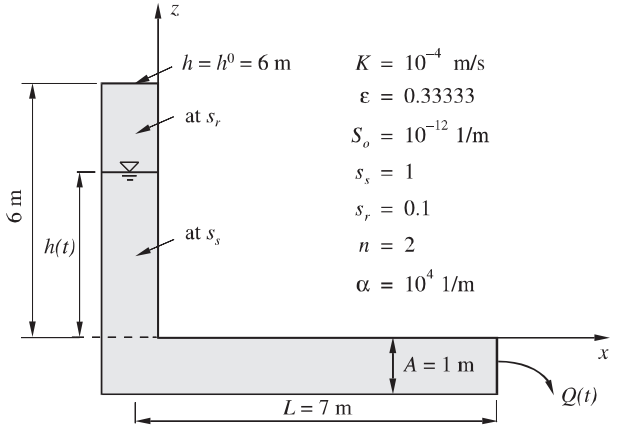


Fig. 10. Sketch of the drainage problem.

(3) and the Brooks-Corey (5) constitutive relationships. The latter offers the advantage to choose the $K_r(s)$ relationship (6) independently of the $\psi(s)$-curve (5).

In this context an analytical expression for the water table descent can be easily derived as

$$\frac{dh}{dt} = -\frac{K}{\varepsilon(s_s - s_r)}\left[\frac{h(t)}{h(t)+L}\right] \tag{60}$$

where $h(t)$ is the water table elevation.

Integrating this equation yields

$$t = \frac{\varepsilon(s_s - s_r)}{K}\left[h^0(1-\varphi) - L\cdot\ln|\varphi|\right], \quad \varphi = \frac{h(t)}{h^0} \tag{61a}$$

and

$$Q(t) = KA\left(\frac{\varphi}{\varphi + L/h^0}\right). \tag{61b}$$

Table 4 lists the analytical results at given relative drawdown $\varphi$. The domain is discretized in 200 quadrilateral finite elements ($\Delta z = 6.5$ cm), where the original problem (Fig. 10) can be modelled by a straight 13-m-long strip. The initial time step is $\Delta t_0 = 10^{-18}$ d. For this example the PCOSN scheme with FE/BE using $\Xi_{max} = 2$ is selected.

Table 4
Analytical results

| $\varphi$ | $t$ (d) | $Q$ (m³/d) | $\int_0^t Q(t)dt$ |
|---|---|---|---|
| 1.0 | 0 | 3.987692 | 0 |
| 0.75 | 0.122006 | 3.380870 | 0.45 |
| 0.50 | 0.272640 | 2.592000 | 0.9 |
| 0.25 | 0.493197 | 1.524706 | 1.35 |
| 0.0372872 | 1.0 | 0.267586 | 1.789 |
| 0 | $\infty$ | 0 | 1.8 |

Initially, the domain is fully saturated at $h^0 = 6$ m and compressibility $S_o$ initiates the drainage process. Using the strong Van Genuchten parameters as stated in Fig. 10 only the variable switching technique was successful while the mixed ($\psi - s$)-form ran into significant convergence difficulties and the standard $\psi$-form even completely failed. The computational results for the PCOSN scheme are listed in Table 5. The agreement with the analytical results (Table 4) is quite good. The solution needs a rather large number of Newton steps (6063 for a simulation time of 1 day with central weighting). However, one can relax (smooth) the problem when setting the parameters equivalent to a free-surface approach [8]. In this case we prefer the Brooks–Corey parametric model (5) and (6) with the following 'simplified' data: $\alpha \approx 1/(\Delta z/2) = 31$ 1/m, $n = 1$, and $\kappa = 1$. The central weighting solution with these Brooks–Corey parameters requires 2544 Newton steps for a 1-day simulation. Note that the reduction of the exponent $\kappa$ to unity is somewhat artificial. However, it is acceptable for this water table problem (see the results presented in Table 5 in comparison to the analytical results of Table 4).

Table 5
Numerical results computed by the PCOSN variable switching technique ($\delta = 5 \cdot 10^{-5}$, central and upstream weighting, FE/BE, $\Xi_{max} = 2$)

| $t$ (d) | Van Genuchten model: $\alpha = 10^4$ 1/m, $n = 2$ | | Brooks–Corey model: $\alpha = 31$ 1/m, $n = 1$, $\kappa = 1$ | | | |
|---|---|---|---|---|---|---|
| | Central weighting | | Central weighting | | Upstream weighting | |
| | $Q$ (m$^3$/d) | $\int_0^t Q(t)dt$ | $Q$ (m$^3$/d) | $\int_0^t Q(t)dt$ | $Q$ (m$^3$/d) | $\int_0^t Q(t)dt$ |
| $10^{-8}$ | 3.9876 | $3.5 \cdot 10^{-8}$ | 3.9618 | $5.2 \cdot 10^{-8}$ | 3.9603 | $4.0 \cdot 10^{-8}$ |
| 0.122006 | 3.3669 | 0.4454 | 3.2917 | 0.4407 | 3.2715 | 0.4394 |
| 0.272640 | 2.6185 | 0.8884 | 2.5026 | 0.8783 | 2.4722 | 0.8722 |
| 0.493197 | 1.5803 | 1.328 | 1.4703 | 1.313 | 1.4434 | 1.300 |
| 1.0 | 0.3285 | 1.727 | 0.2742 | 1.686 | 0.2679 | 1.665 |

The upstream weighting was not successful for the Van Genuchten model. Applying the Brooks–Corey model with central and upstream weighting gave comparable results as listed in Table 5. The number of Newton steps slightly increased to 2818 for a 1-day simulation if upstream weighting was applied.

In estimating the TMBE ($T = 1$ d) error (54) we found $O(10^{-2})$ for both the Van Genuchten model and the Brooks–Corey model. This estimate is conform to the mass defects which are detected in the comparisons of the numerical results of Table 5 to the analytical results of Table 4.

### 8.3. Perched water table problem

Kirkland et al. [26] presented a two-dimensional problem of a developing perched water table surrounded by very dry unsaturated conditions. It is a good test problem to show the variable switching ability in both unsaturated and saturated zones. The problem is described in Fig. 11. Water infiltrates with a very large rate into a dry soil at $\psi^0 = -500$ m and encounters a clay barrier which allows for the formation of a perched water table. All boundaries are no flow except where the infiltration is imposed. The material properties of the problem are summarized in Table 6 for the Van Genuchten–Mualem parametric model. Both the PCOSN and the TBFN scheme are used with $\delta = 10^{-4}$ and $\Delta t_0 = 10^{-5}$ d. Additionally, TBFN is constrained by $\Xi_{max} = 2$. The symmetric half of the domain is discretized in a $50 \times 60$ quadrilateral mesh (3111 nodes) according to the spatial discretization used by Kirkland et al. [26] and Forsyth et al. [13].

A comparison of the pressure contours at 1-day with Kirkland et al.'s results reveals an acceptable agreement as displayed in Fig. 12. The zero pressure contours agree quite well while the −4000 kPa isobar equivalent of Kirkland et al.'s results is slightly ahead, forming a more diffusive vertical pressure front compared to the present solution. The higher sharpness of the present profile is also identified in comparison to Forsyth et al.'s saturation contours (Fig. 13). Forsyth et al. [13] used an aggressive target-based time marching scheme similar to

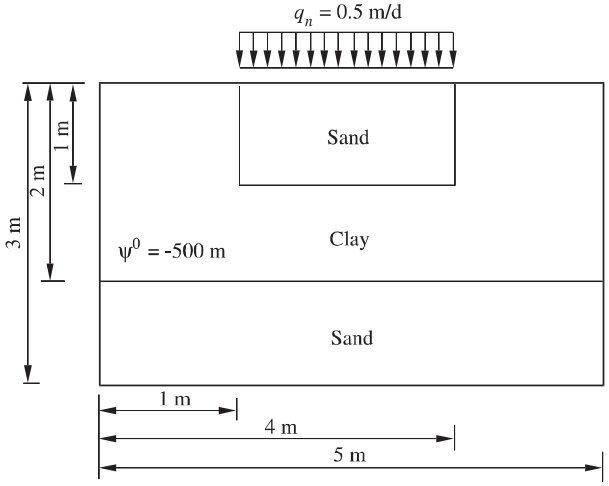


Fig. 11. Perched water table problem (modified from [26]).

Table 6
Material properties for the perched water table problem

| Material | $K$ (m/s) | $\varepsilon$ (1) | $s_r$ (1) | $\alpha$ (1/m) | $n$ (1) |
|---|---|---|---|---|---|
| Sand | $6.262 \cdot 10^{-5}$ | 0.3658 | 0.07818 | 2.80 | 2.2390 |
| Clay | $1.516 \cdot 10^{-6}$ | 0.4686 | 0.2262 | 1.04 | 1.3954 |

the present TBFN method and got the solution after 120 Newton steps. The present PCOSN and TBFN schemes needed many more steps with the given control parameters. This is probably due to a lack of smoothness in the parametric curves near full saturation. The variable switching technique for the PCOSN (FE/BE) technique at central weighting required 1211 time steps and 1556 Newton steps, meaning that about 30% of the steps had to be rejected and repeated. In contrast, the TBFN scheme became less efficient. Only 582 time steps were needed but the total number of Newton iterations increased to 3381 steps. Similar results were found for upstream weighting. Pressure and saturation profiles are given in Figs. 12 and 13, respectively.

As displayed in the time step histories for both schemes in Fig. 14 the TBFN scheme progresses faster at the beginning, while the PCOSN scheme takes smaller step sizes due to the temporal discretization

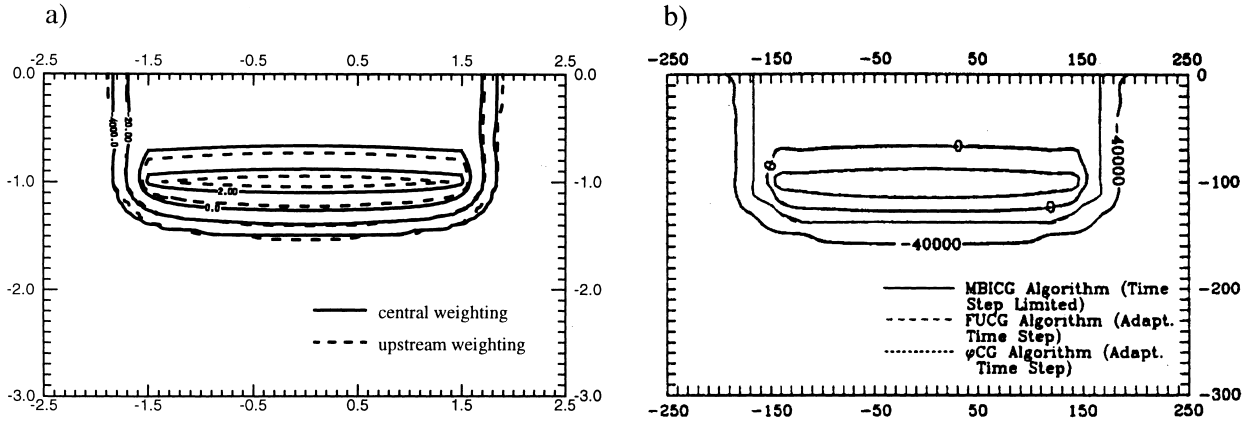


Fig. 12. Simulated pressure contours at $t = 1$ d: (a) present results, PCOSN and TBFN, FE/BE, central and upstream weighting, pressure contours in (kPa), lengths in (m); (b) Kirkland et al.'s results [26], pressure head contours in (cm), lengths in (cm).

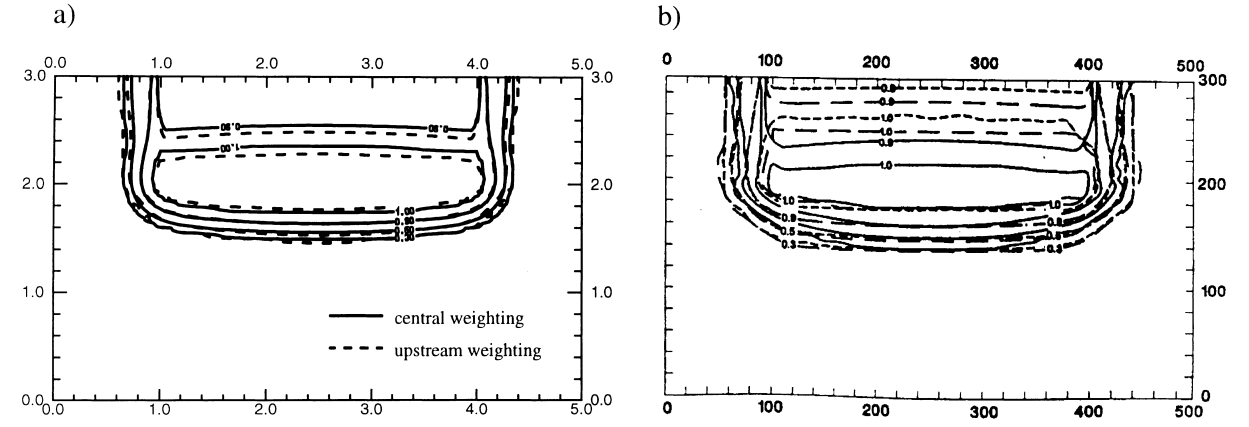


Fig. 13. Simulated saturation contours at $t = 1$ d: (a) present results, PCOSN and TBFN, FE/BE, central and upstream weighting, lengths in (m); (b) Forsyth et al.'s results [13]; (- - -) one phase, upstream weighting; (· · ·) one phase, central weighting; (——) two phases, upstream weighting, lengths in (cm).

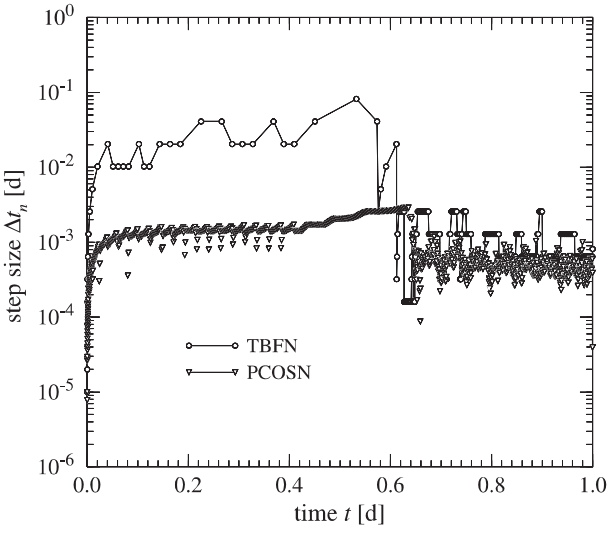


Fig. 14. Time step histories of the perched water table problem for the TBFN and PCOSN schemes (FE/BE, central weighting) using $\delta = 10^{-4}$ and $\Delta t_0 = 10^{-5}$ d ($\Xi_{max} = 2$ for TBFN); required time steps: 582 (Newton 3381) for TBFN and 1211 (Newton 1556) for PCOSN.

accuracy requirements. As soon as the perched water table is formed (nodes become saturated) the convergence criterion of the TBFN scheme forces smaller steps. The aggressive selection strategy leads to a rapid growth of the provisional time step size. However, the latter is invariably too large for the convergence of Newton iterations and the larger step sizes have to be discarded. Oscillations in the step size result in the poor performance of the TBFN scheme for the present problem, whereas the PCOSN solution strategy is not affected by such oscillations. Apparently, the TBFN strategy can be improved by refining the time stepping control (e.g., introducing a multiple set of decision parameters). To this end, Forsyth and Simpson [12] proposed a manual monitoring via a file-based checking procedure.

The simulations with the PCOSN and TBFN schemes give identical results (Figs. 12 and 13) because the required step number is sufficiently high and meets the accuracy requirements. Considering the results found in the above sections, the differences between the present

and Kirkland et al.'s as well as Forsyth et al.'s results can mainly be attributed to temporal discretization effects. Typically, a smaller step number generates a phase lead and a smoother front. This will be also confirmed in the following examples.

The TMBE ($T = 1$ d) balance error (54) was found to be of O($10^{-4}$) for the PCOSN and of O($10^{-5}$) for the TBFN scheme.

### 8.4. Infiltration in a large caisson

#### 8.4.1. Forsyth et al.'s problem

The infiltration process in a large caisson consisting of heterogeneous materials at dry initial conditions has been thoroughly studied by Forsyth et al. [13]. We choose this problem to show the power of the variable switching technique and to identify solution differences caused by the time stepping and iteration control alternatives. Fig. 15 presents a schematic view of the 2D cross-sectional problem. All boundaries are impervious except the infiltration boundary section on top. Two initial pressure head conditions of $\psi^0 = -7.34$ m and $\psi^0 = -100$ m are simulated. Table 7 lists the material properties used for the different zones of the domain. Both the PCOSN and the TBFN schemes are applied with $\delta = 10^{-4}$, $\Delta t_0 = 10^{-3}$ d (TBFN is again constrained by $\Xi_{\max} = 2$) with central and upstream weighting. Fully implicit FE/BE strategies are selected. The spatial discretization is $89 \times 20$ quadrilateral elements (1890 nodes) as in Forsyth et al. [13].

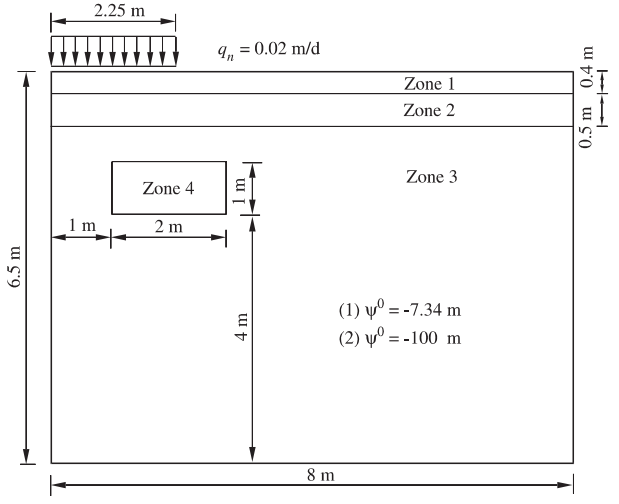


Fig. 15. Forsyth et al.'s infiltration problem (modified from [13]).

Table 7
Material properties for Forsyth et al.'s problem (Van Genuchten–Mualem parametric model)

| Zone | $K$ (m/s) | $\varepsilon$ (1) | $s_r$ (1) | $\alpha$ (1/m) | $n$ (1) |
|---|---|---|---|---|---|
| 1 | $9.153 \cdot 10^{-5}$ | 0.3680 | 0.2771 | 3.34 | 1.982 |
| 2 | $5.445 \cdot 10^{-5}$ | 0.3510 | 0.2806 | 3.63 | 1.632 |
| 3 | $4.805 \cdot 10^{-5}$ | 0.3250 | 0.2643 | 3.45 | 1.573 |
| 4 | $4.805 \cdot 10^{-4}$ | 0.3250 | 0.2643 | 3.45 | 1.573 |

Based on the given control parameters the TBFN scheme was about four times faster than the PCOSN scheme as indicated in Table 8. On the average 3–4 Newton steps were required for the TBFN strategy at each time step. The PCOSN scheme provided a quite perfect time stepping control without repeated time steps. The extra costs for the PCOSN scheme are reflected by an increased temporal accuracy, as required by the error control. The results at 30 days can be seen in Figs. 16 and 17 for $\psi^0 = -7.34$ m and $\psi^0 = -100$ m, respectively, in comparison to Forsyth et al.'s findings [13].

Surprisingly, the PCOSN results are rather depart from the TBFN results, especially for the case $\psi^0 = -7.34$ m. The saturation front is significantly diffused by the 'low-cost' TBFN simulation using the 'aggressive' control parameters (59) while the PCOSN provides a much steeper saturation profile. Expectedly, Forsyth et al.'s results [13] agree quite well with the poorer TBFN solutions since they performed an even smaller number of Newton steps (29 steps at $\psi^0 = -7.34$ m and 48 steps at $\psi^0 = -100$ m, for central weighting). This example clearly illustrates how far a seemingly accurate, convergent and efficient solution can be from a more accurate prediction independent of the use of central and upstream weighting. Control parameters smaller than (59) have to be chosen for the TBFN to enforce smaller time step sizes and to find results comparable to the PCOSN.

It is apparent that the present problem is sensitive to discretization errors. The influence of the spatial discretization is illustrated in Fig. 18 for the case $\psi^0 = -100$ m. The results of structured coarse meshes ($90 \times 21$ and $21 \times 90$ nodes) are compared to a dense unstructured mesh consisting of 56 960 triangular elements (28 917 nodes). This dense mesh is generated by

Table 8
Solution effort for Forsyth et al.'s problem (FE/BE)

| | $\psi^0 = -7.34$ m | | | | $\psi^0 = -100$ m | | | |
|---|---|---|---|---|---|---|---|---|
| | PCOSN | | TBFN | | PCOSN | | TBFN | |
| | central | up stream | central | up stream | central | up stream | central | up stream |
| Time steps | 199 | 174 | 15 | 15 | 279 | 251 | 16 | 15 |
| Total Newton steps | 200 | 174 | 51 | 67 | 279 | 251 | 69 | 69 |

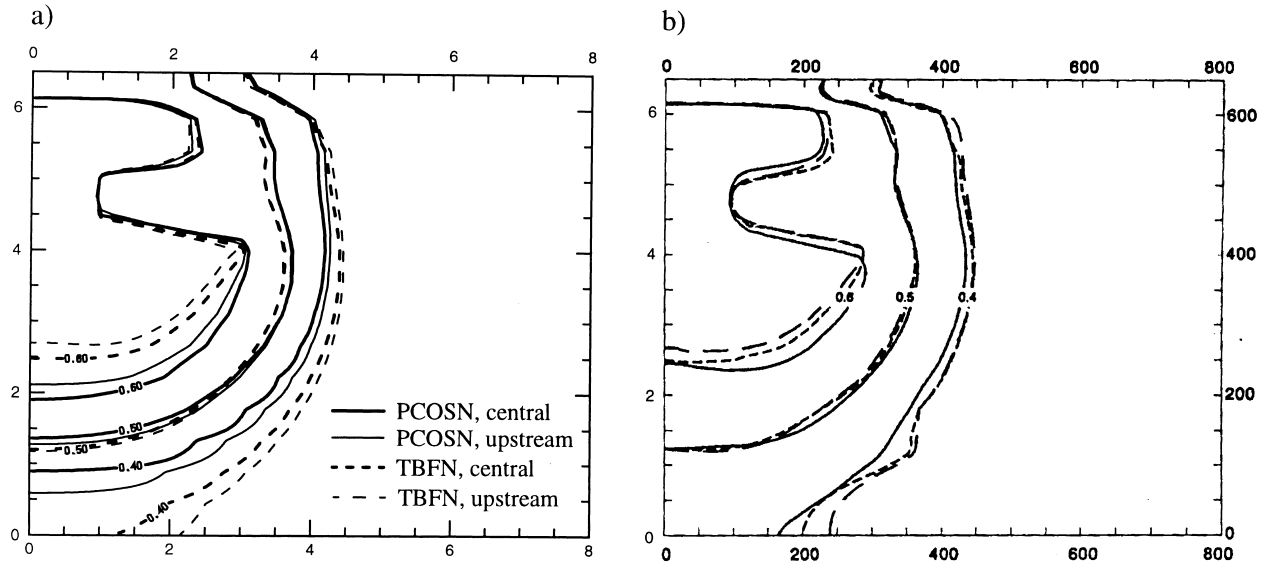


Fig. 16. Computed saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -7.34$ m: (a) present solutions by PCOSN and TBFN, central and upstream weighting, lengths in (m); (b) Forsyth et al.'s results [13]; (- - -) one phase, upstream weighting; (· · ·) one phase, central weighting; ( — ) two phases, upstream weighting, lengths in (cm).

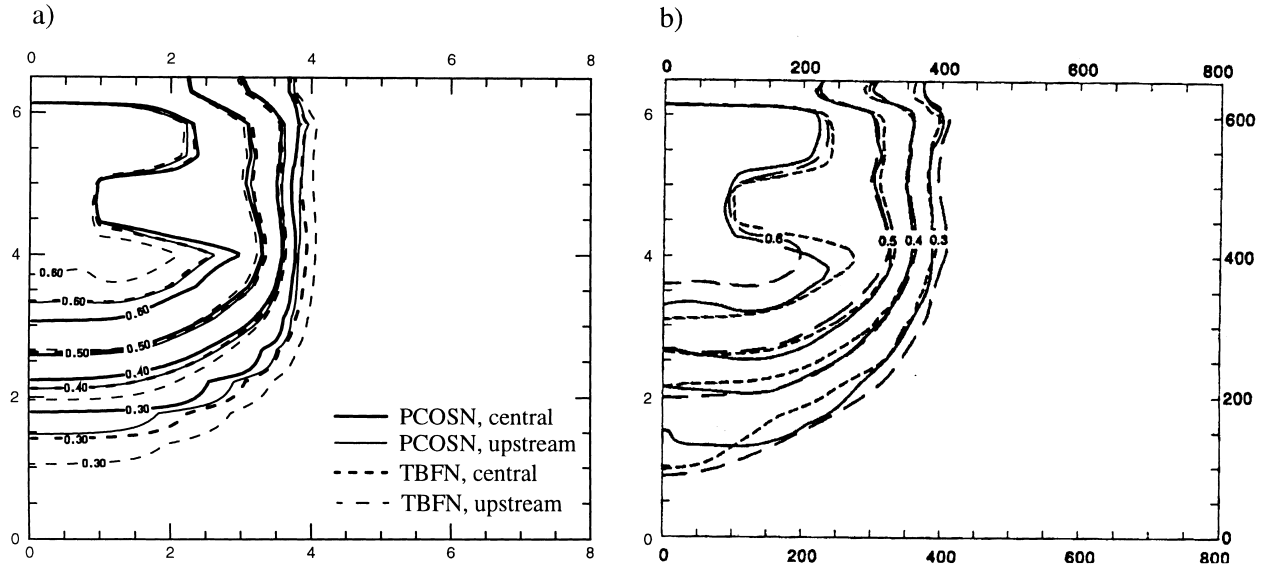


Fig. 17. Computed saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -100$ m: (a) present solutions by PCOSN and TBFN, central and upstream weighting, lengths in (m); (b) Forsyth et al.'s results [13]; (- - -) one phase, upstream weighting; (· · ·) one phase, central weighting; ( — ) two phases, upstream weighting, lengths in (cm).

splitting each quadrilateral into two triangles followed by a double total refinement into four triangles ($20 \times 89 \times 2 \times 4 \times 4$). It shows how a coarse meshing in one direction can lead to phase lag errors and smearing of the saturation profiles.

The history of the residual error $\|R\|_{L_2}$ depending on the selected time stepping schemes and the initial pressure head $\psi^0$ is plotted in Fig. 19. The one-step Newton scheme (PCOSN) terminates with errors of $O(10^{-5})$ while the TBFN is, at least, one order better. This naturally results from the full Newton technique incorporated in the TBFN, where, at least, two iteration steps are performed and convergence in the residuals $\|R\|_{L_2}$ is quadratic. Accordingly, we estimate a TMBE ($T = 30$ d) of $O(10^{-5})$ for the PCOSN and of $O(10^{-6})$ for the TBFN.

### *8.4.2. Forsyth and Kropinski's problem*

Forsyth and Kropinski [14] modified the above infiltration problem of Fig. 15 by increasing the pore size distribution index $n$ to 5 for the zones 3 and 4. The other parameters remain unchanged and correspond to

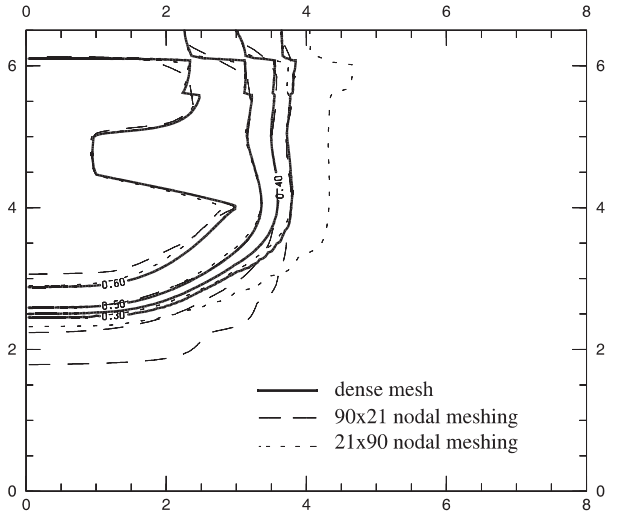


Fig. 18. Influence of spatial discretization, computed saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -100$ m: dense mesh consists of 56,960 triangles and 28,917 nodes, central weighting, PCOSN (FE/BE) scheme with 2507 implicit time steps and 3596 Newton steps, lengths in (m).

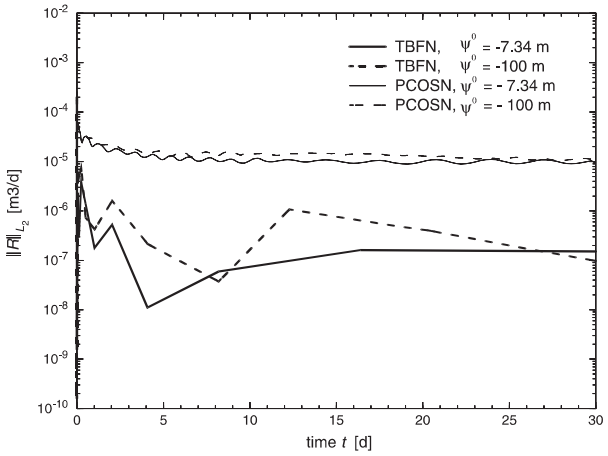


Fig. 19. History of residual error $\|R\|_{L_2}$ for the TBFN and PCOSN schemes with $\delta = 10^{-4}$, RMS error convergence criterion (43), central weighting and $90 \times 21$ mesh.

Table 7. This increase of $n$ makes the capillary pressure curve very flat at intermediate saturation values and spurious local maxima and minima can result for coarse meshes. This is shown in Fig. 20 for a structured $90 \times 21$ nodal meshing and a central weighting. The comparison with Forsyth and Kropinski [14] indicates mesh effects. Although using the same mesh, differences at material interfaces and at the bottom of the caisson are detected. These may result from different nodal spacing at these locations. The PCOSN required 1202 time steps with 2015 Newton steps, whereas the TBFN only took 146 time steps and 809 Newton iterations. As shown in Fig. 20 the reduced stepping by TBFN leads to smearing and phase lead errors, however, only for the advanced saturation contours while the remaining part is close to the PCOSN results.

Upstream weighting can be used to damp out the spurious oscillations in the saturation distributions. Fig. 21 compares the present upstream solution with Forsyth and Kropinski's result. The agreement is quite good. Both upstream techniques damp out the wiggles appearing in the central weighting solutions (Fig. 20). Differences in the lag of the saturation profile are probably due to the different nodal spacing used in the present and Forsyth and Kropinski's [14] solutions.

A more appropriate meshing of the problem (i.e., $21 \times 90$ instead of $90 \times 21$) can considerably improve

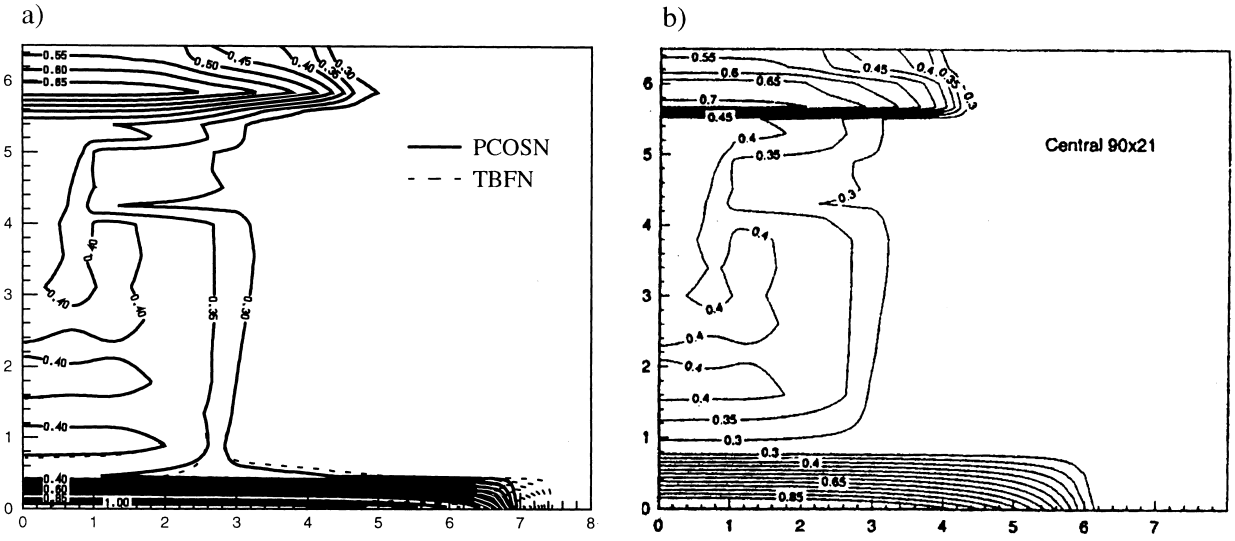


Fig. 20. Saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -100$ m, and central weighting: (a) present solutions by PCOSN and TBFN, $90 \times 21$ nodal meshing; (b) Forsyth and Kropinski's results [14]; lengths in (m).

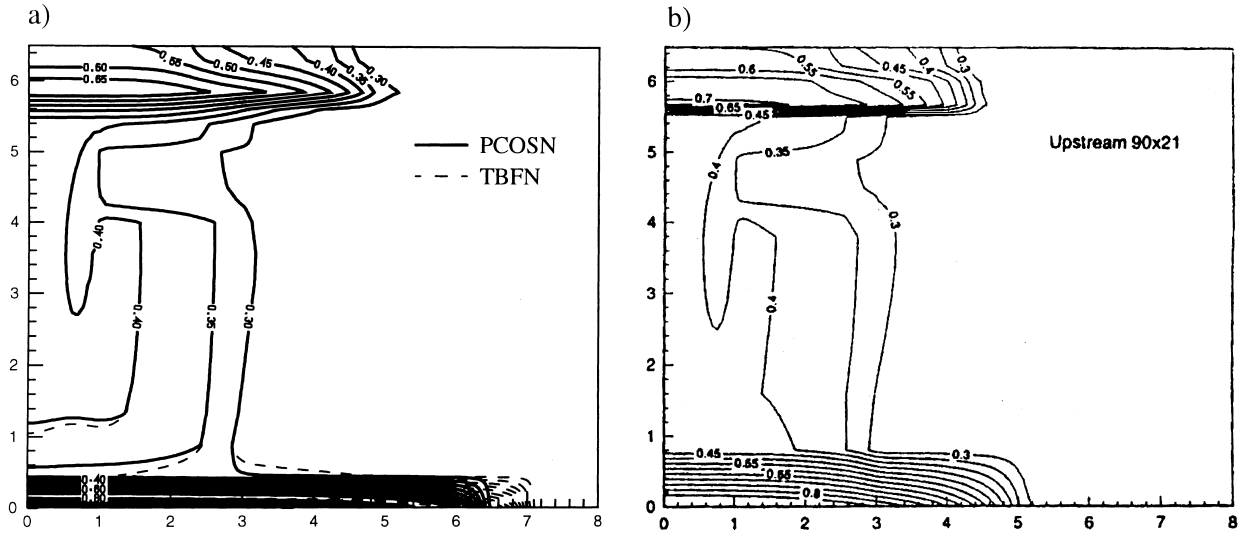


Fig. 21. Saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -100$ m, and upstream weighting: (a) present solutions by PCOSN and TBFN, $90 \times 21$ nodal meshing; (b) Forsyth and Kropinski's results [14]; lengths in (m).

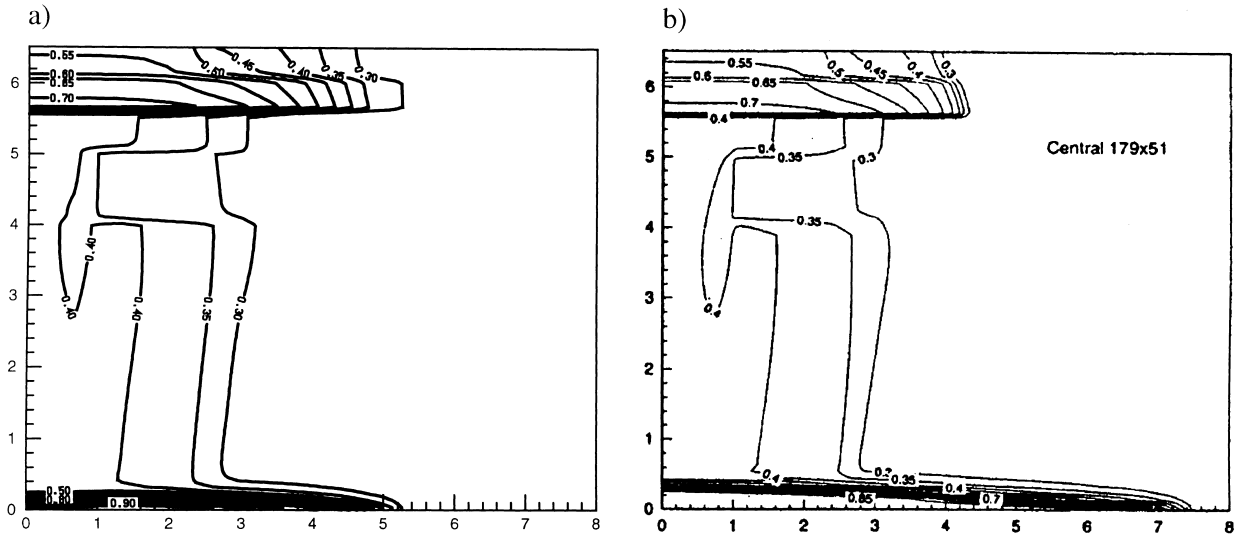


Fig. 22. Saturation contours for refined meshes at $t = 30$ d, initial pressure head $\psi^0 = -100$ m, and central weighting: (a) present solutions by PCOSN, $21 \times 90$ nodal meshing; (b) Forsyth and Kropinski's results [14]; lengths in (m).

the results (Fig. 22(a) and (b)). The solution can be compared to the results obtained with the dense mesh (28.917 nodes) shown in Fig. 23. Sharper saturation contours occur at the material interfaces. The medium becomes fully saturated at the bottom of the caisson forming a typical saturation 'tongue'. Its size is quite sensitive to spatial and temporal discretizations as revealed by the comparison to Fig. 22. In contrast, Forsyth and Kropinski [14] predict a lead in the saturation pattern (Fig. 22(b)).

In checking the mass balance errors TMBE ($T$), Eq. (54), we estimate the same order as indicated in the above problem of Section 8.4.1: TMBE ($T = 30$ d) of $O(10^{-5})$ for the PCOSN and of $O(10^{-6})$ for the TBFN.

### 8.5. Capillary barrier modeling

In unsaturated flow conditions a capillary barrier often appears at the contact of a layer of fine soil overlying a layer of coarse soil [33,47]. If the layer interface is tilted, moisture infiltrating in the fine layer will be diverted and flow down the contact. In practical applications, a capillary barrier can be built by placing a fine layer (e.g., fine sand) over an inclined coarse layer (e.g., gravel). To simulate capillary barriers numerical schemes have to tackle large parameter contrasts, highly exaggerated and distorted geometries as well as very dry initial conditions. Focusing on steady-state solutions, which are of the most practical interest here, and assuming that there is no bifurcation in the development

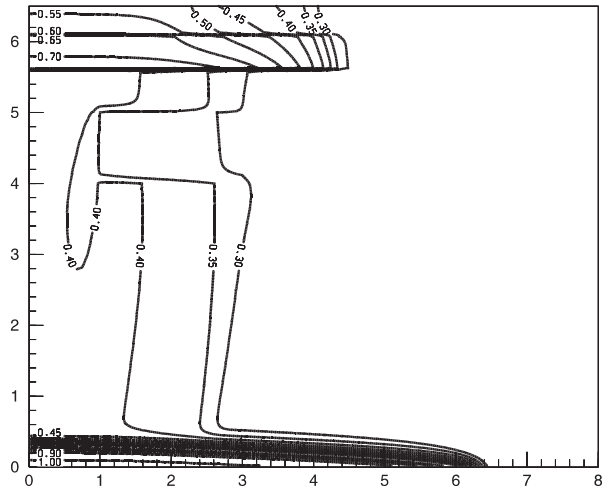


Fig. 23. Present saturation contours for the dense mesh (28,917 nodes) at $t = 30$ d, initial pressure head $\psi^0 = -100$ m, central weighting, lengths in (m).

of the capillary diversion, the TBFN scheme seems be the most effective solution technique for this class of problems.

### 8.5.1. Webb's problem

Oldenburg and Pruess [33] presented a first numerical study of a 2D tilted capillary barrier. To find reasonable results they introduced an upstream weighting method. However, both from the qualitative and quantitative point of view their results became generally poor and no agreement with analytical results [39] could be achieved. More recently, Webb [47] could improve the steady-state results by using an upstream weighting technique agreeing well with Ross' analytical prediction [39]. We use Webb's capillary barrier problem [47] to study the capability of the variable switching technique for both central and upstream weighting.

Webb's capillary barrier consists of a two (fine over coarse) layer configuration with a total thickness of 1 m. The fine and coarse layers are both 0.5 m thick, and the dip of the layers is 5% (2.86°). The parameters of the two layers are summarized in Table 9. The infiltration rate at the surface of the domain is 0.0048 m/d. The left boundary is impervious and the right and bottom boundaries allow for drainage. This can be done in several ways. We attempted different alternatives: constrained point sinks, gradient-type boundary conditions and potential-type boundary conditions. In consideration of the extreme parameter situation of the fine and coarse layers (*cf.* Table 9) we found a better convergence behavior for potential-type boundary conditions, where the hydraulic head $h$ is imposed. Since the $\alpha$-parameter of the coarse layer is very large the influence of the location of the water table (the $\psi = 0$ condition) cannot be significant. It is thus sufficient to set the water table at the right lower corner of the domain (at $z = 0$) and prescribe a $h = 0$ Dirichlet boundary condition along the bottom and the right boundaries. In accordance with these boundary conditions a corresponding hydrostatic initial condition is assumed, i.e., a vertical linear distribution of $h^0$ in the range from 0 to $-6$ m. This results in averaged initial saturations $s^0$ of 0.394872 for the fine layer and 0.02864 for the coarse layer which is very close to the residual saturations $s_r$ (*cf.* Table 9). The model domain is appropriately discretized in quadrilateral elements as displayed in Fig. 24. At the layer contact the element thickness is 0.005 m, and gradually increases with the distance from the interface. The implicit time stepping (FE/BE) was used with $\Delta t_0 = 10^{-3}$ d.

Table 9
Material properties for Webb's capillary barrier problem (Van Genuchten–Mualem parametric model)

| Parameter | Upper layer (fine) | Lower layer (coarse) |
|---|---|---|
| $\varepsilon$ (1) | 0.39 | 0.42 |
| $K$ (m/s) | $2.1 \cdot 10^{-4}$ | 0.1 |
| $s_r$ (1) | 0.394872 | 0.028571 |
| $s_s$ (1) | 1.0 | 1.0 |
| $n$ (1) | 5.74 | 2.19 |
| $\alpha$ (1/m) | 3.9 | 490.0 |

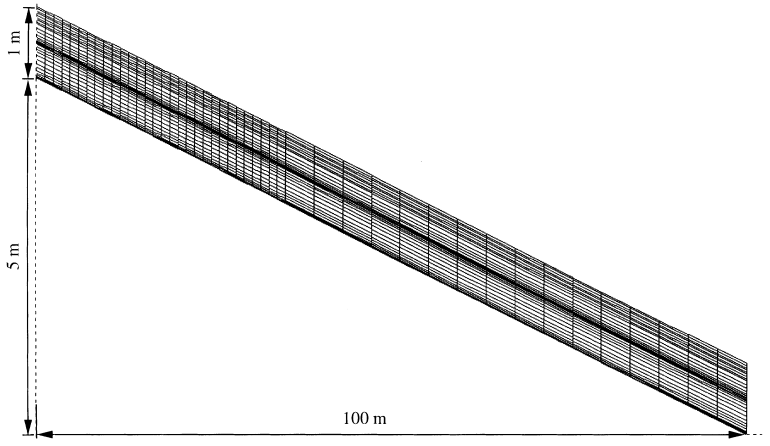


Fig. 24. Model domain and mesh (1472 quadrilaterals with 1551 nodes) for Webb's capillary barrier problem [47] (exaggeration 10:1).

Surprisingly, the TBFN scheme ran into significant convergence difficulties. The reason is that a fully saturated zone is quickly formed in the upper layer along the material interface. Such a situation is similar to the perched water problem previously studied in Section 8.3 where the PCOSN scheme became superior to the TBFN. For the present problem successful solutions were obtained by PCOSN running over a time period of 100 days. At this time, the flow budget has reached equilibrium and the capillary diversion effect has settled down. Due to the sharp parameter contrasts we select for this task the maximum error norm (44) instead of the integral RMS norm (43). Here, an error tolerance of $\delta = 10^{-3}$ turned out to be sufficient.

Fig. 25 exhibits the computed saturation distribution at 100 days. It reveals how the saturated zone has built up along the contact zone in the fine layer while the saturation in the coarse layer remains only slightly above the residual saturation. From such a saturation pattern the capillary diversion cannot be identified. However, the integration of the velocity field in form of streamlines clearly illustrates the capillary diversion effects, as shown in Fig. 25. The diversion is maintained up to a certain distance, the diversion length, past which an amount of water equal to the infiltration rate enters the coarse layer.

A comparison of the above results with Ross' analytical formula [39] and the numerical results obtained by Webb [47] can be expressed as a function of the leakage/infiltration ratio. The theoretical value of the diversion length determined from Ross' formula is 32.6 m for the present parameters (note, Webb [47] computed 33.2 m). As evidenced in Fig. 26 there is a good qualitative and quantitative agreement between the analytical and the numerical results. Note here that Webb's solution is based on an upstream weighting scheme. The present method was able to find solutions for both central and upstream weighting. As seen in Fig. 26 the differences between upstream and central weighting are relatively small. Upstream weighting damps the slight oscillations of the downstream velocity field. The breakthrough point is not significantly affected.

It should be mentioned that the specific advantages of the variable switching technique disappear in the present capillary barrier problem. Since the initial pressures remain moderate and since conservation properties do not play a role for computing a steady-state solution, the

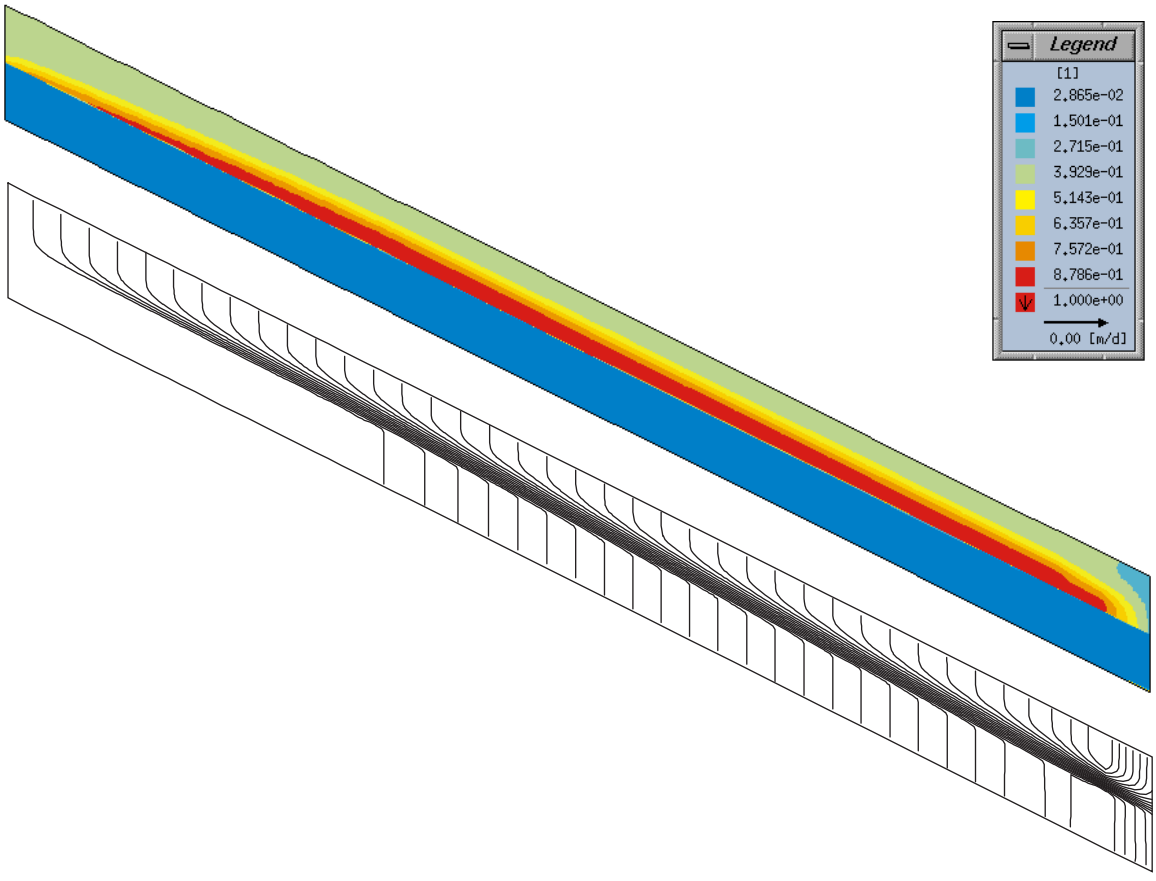


Fig. 25. Computed saturation and streamline patterns for Webb's capillary barrier [47] (exaggeration 10 : 1).

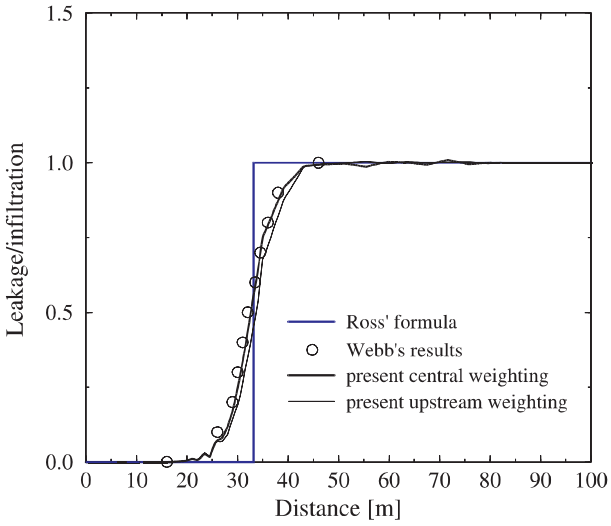


Fig. 26. Leakage/infiltration ratio in the coarse layer for both central and upstream weighting compared to Ross' analytical formula [39] and Webb's numerical results [47].

classic $\psi$-based form becomes an effective alternative. We confirmed the above solutions for the $\psi$-model (32), using the predictor–corrector time stepping scheme for both FE/BE and AB/TR, and without the Newton method.

*8.5.2. Forsyth and Kropinski's problem*

A different capillary barrier problem has been recently considered by Forsyth and Kropinski [14]. The problem is described in Fig. 27. The material properties and the initial pressure conditions for the different layers are given in Table 10. As indicated the initial conditions are very dry. The infiltration rate at the surface of the cross-sectional domain is 15 cm/yr. The mesh is shown in Fig. 27. It consists of 5002 quadrilateral linear elements with 5146 nodes. As seen, the element size is highly variable in the vertical direction. At the sand-gravel interface the elements have a thickness as small as 0.002 m. The left and right vertical boundaries are considered impervious. To model free drainage at the bottom of the domain the gradient-type boundary condition $q_n^{h\nabla} = K|_{\text{bottom}} = 0.23985\ \text{m/d}$ applies there.

We used both the PCOSN and the TBFN scheme with $\delta = 10^{-4}$, $\Delta t_0 = 10^{-5}$ d and $\Xi_{\max} = 5$. Due to the extremely dry initial conditions the PCOSN scheme required an unacceptable number of time steps. On the other hand, the TBFN scheme, not constrained by temporal discretization error bounds, provided solutions with a much smaller number of time steps (and Newton steps).

We ran the problem for a simulation time of 30 yr with the TBFN and applying both the $L_2$ (43) and $L_\infty$

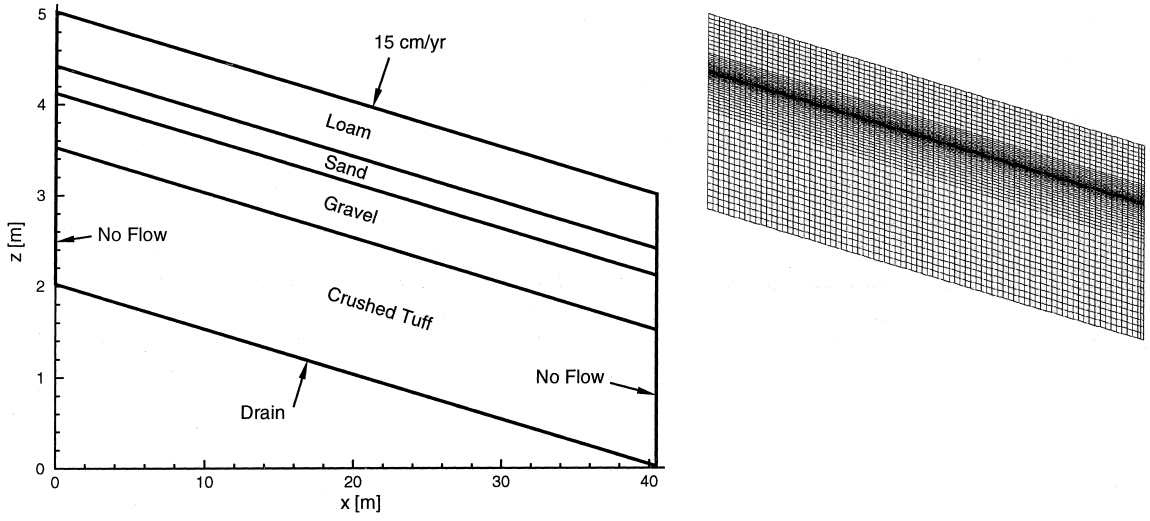


Fig. 27. Capillary barrier model domain (modified from [14]) and mesh (5002 isoparametric bilinear elements with 5146 nodes).

Table 10
Material properties and initial pressure for Forsyth and Kropinski's capillary barrier problem (Van Genuchten–Mualem parametric model)

| Zone | $K$ (m/s) | $\varepsilon$ (1) | $s_r$ (1) | $\alpha$ (1/m) | $n$ (1) | $\psi^0$ (kPa) |
|---|---|---|---|---|---|---|
| Loam | $1.668 \cdot 10^{-5}$ | 0.452 | 0.0752 | 4.3 | 1.246 | $-10^6$ |
| Sand | $6.573 \cdot 10^{-5}$ | 0.345 | 0.046 | 6.34 | 1.53 | $-10^6$ |
| Gravel | $3.502 \cdot 10^{-3}$ | 0.419 | 0.074 | 469.0 | 2.57 | $-30$ |
| Crushed tuff | $2.776 \cdot 10^{-6}$ | 0.345 | 0.032 | 1.43 | 1.506 | $-6 \cdot 10^{10}$ |

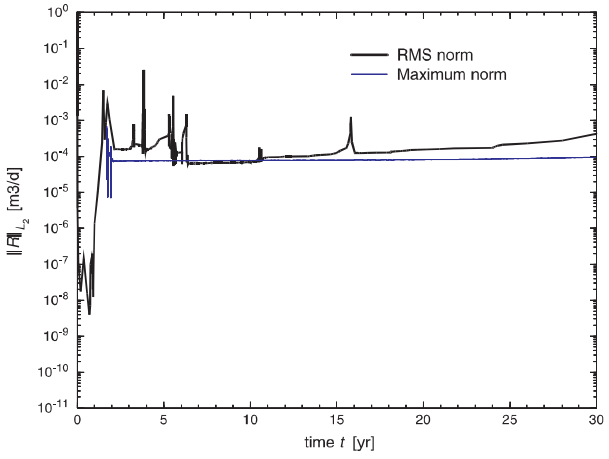


Fig. 28. History of residual error $\|R\|_{L_2}$ for the RMS $L_2$ and maximum $L_\infty$ convergence criteria with $\delta = 10^{-4}$ and central weighting.

(44) error norms for terminating the Newton iteration. The evolution of the residual error $\|R\|_{L_2}$ for both norms is depicted in Fig. 28. It reveals that the $L_2$ criterion produces residuals in the range $10^{-3}$–$10^{-2}$ ($m^3$/d). For this case the integral total mass balance error is TMBE $(T = 30\text{ yr}) \approx 1.2 \times 10^{-2}$, which cannot be tolerated. The results for the $L_\infty$ criterion is better by about one order (*cf.* Fig. 28) and gives TMBE $(T = 30\text{ yr}) \approx 4.7 \times 10^{-3}$. Accordingly, only the results obtained under the $L_\infty$ criterion will be discussed.

The 30-year simulation under the $L_\infty$ convergence criterion took about 5000 time steps (with about $10^4$ total Newton steps) for both the upstream and the central weighting. We found the solutions in form of saturation and streamline patterns as displayed in Fig. 29(a), Fig. 30(a) and Fig. 31.

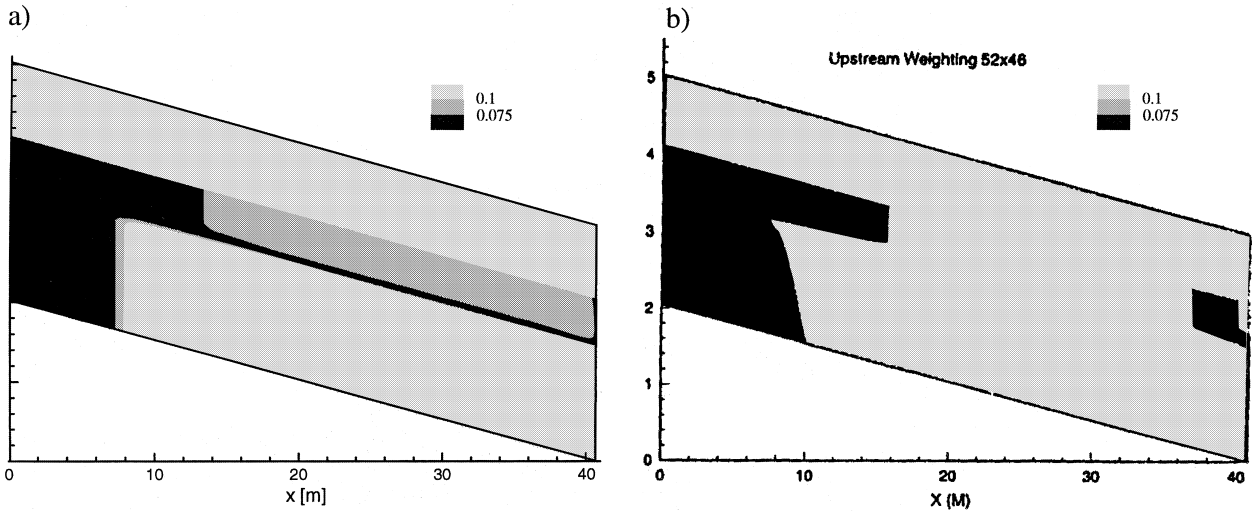


Fig. 29. Simulated saturation patterns at $t = 30$ yr: (a) present solution by TBFN and upstream weighting; (b) upstream weighting solution obtained by Forsyth and Kropinski [14].

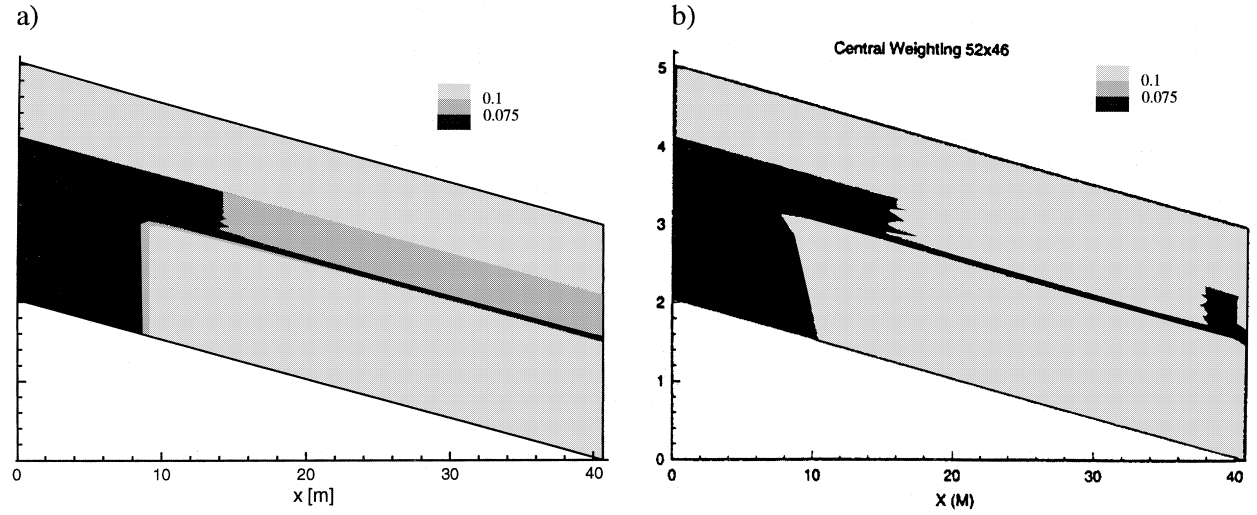


Fig. 30. Simulated saturation patterns at $t = 30$ yr: (a) present solution by TBFN and central weighting; (b) central weighting solution obtained by Forsyth and Kropinski [14].

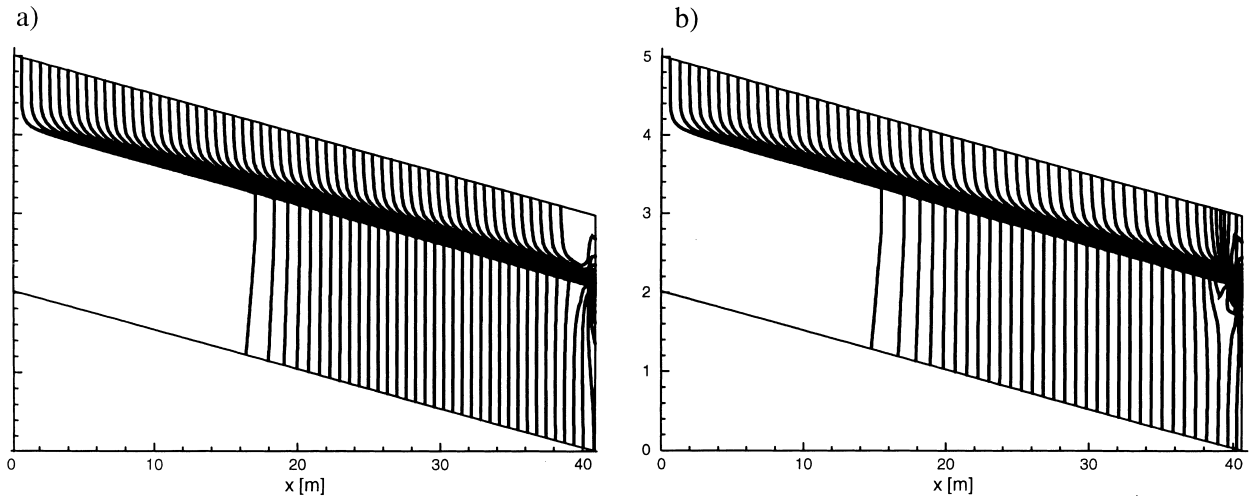


Fig. 31. Simulated streamline patterns at $t = 30$ yr, TBFN for (a) central and (b) upstream weighting.

Forsyth and Kropinski [14] used both central and upstream weighting at two grid resolutions ($52 \times 46$ and $103 \times 92$). They predict that the capillary barrier fails with a diversion length of about 10 m characterized by a saturation distribution as exemplified in Fig. 29(b) for upstream weighting and Fig. 30(b) for central weighting with the $52 \times 46$ grid.

The present simulations confirm Forsyth and Kropinski's results [14]. The computed saturation distributions are displayed for three specific contour levels in Fig. 29(a) for the upstream weighting and in Fig. 30(a) for the central weighting. Some details are depart from Forsyth and Kropinski's simulations. It can be assumed that most of them is caused by different boundary conditions. Forsyth and Kropinski imposed a seepage point on the right-hand side boundary and handled the bottom of the tuff layer as a no-flow boundary, however, at a far vertical position. In the present model, such a seepage point is not imposed and the bottom of the tuff is fully handled as a free-drain boundary at the actual position as shown in Fig. 27. For the central weighting (Fig. 30(a)) we note a jagged saturation profile which disappears for upstream weighting (Fig. 29(a)). A small strip of lower saturation can be seen along the gravel-tuff interface in both the upstream and the central solutions. Forsyth and Kropinski found it only in their central weighting solution (Fig. 30(b)).

The streamline patterns in Fig. 31 illustrate the effect of the capillary barrier at the sandgravel material interface. Only slight differences exist between upstream and central weighting. The streamlines reveal that the diversion length is obviously somewhat larger than 10 m. Actually, the velocity distribution along the bottom of the tuff layer indicate a leakage increase from zero at about 10 m to the infiltration rate at about 25 m, as depicted in Fig. 32. The relatively smooth breakthrough results from the complex layered structure of this capillary barrier. The breakthrough curve is slightly ahead for the upstream weighting. An evaluation of Ross' analytical formula [39] using the above Van Genuchten parameter for the sand and gravel zones (Table 10) gives a diversion length of 17.9 m. This value is in good agreement with the present numerical simulations as seen in Fig. 32.

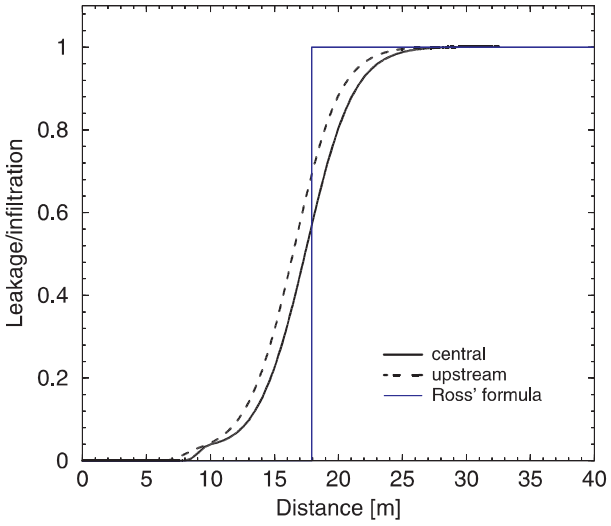


Fig. 32. Leakage/infiltration ratio in the tuff layer.

## 9. Closure

The primary variable switching technique has proved to be a powerful and cost-effective solution strategy for unsaturated flow problems. Compared to conventional approaches based on the $\psi$-form and the mixed ($\psi - s$-) form of the Richards equation, with either Picard or

Newton iteration, the primary variable switching technique can reduce the solution effort by orders. More specifically, for very dry initial conditions, the primary variable switching technique appears as the only practical way to get reasonable solutions. This has been shown in a number of difficult examples. The advantages of the primary variable switching technique can be summarized by the following items. It is

- unconditionally mass-conservative,
- very effective and robust for dry initial conditions,
- a Newton-based iteration method with quadratic convergence representing a 'natural' approach for the approximation of highly nonlinear problems combined to constrained relationships (primary and secondary variables), and
- a general analysis method suitable for single and multi-phase flow problems.

The price to be paid for the primary variable switching technique is in assembling and solving the unsymmetric equation system at each time and Newton step. For unsaturated flow the Jacobian can easily be constructed in an analytical manner to reduce the computational effort. For the most cases studied, however, the increased effort in handling the unsymmetric system is largely compensated by the fast convergence behavior.

Nevertheless, we do not claim to have a panacea for all variably saturated flow problems. We presented a wide spectrum of examples to benchmark the technique and compare our results with previous findings. We found some differences. First of all, the iterative solution procedure embedded in the primary variable switching technique have proved to be of prime importance. We studied both a temporally error-controlled predictor–corrector one-step Newton scheme (PCOSN) and a commonly used [12–14,24] target-based full Newton scheme (TBFN). While the PCOSN satisfies a temporal discretization error at each time (and iteration) step, the TBFN is controlled by the Newton convergence criterion only and does not necessarily satisfy a discretization error. As a result, the PCOSN and the TBFN schemes can provide different solution behaviors. Roughly speaking, the PCOSN needs often more steps, however, gives more accurate solutions. Its numerical control is much simpler for practical use. Only one control parameter, the error tolerance, has to be specified. On the other hand, the TBFN often requires a smaller (sometimes a significantly reduced) number of steps to accomplish a simulation time. In analyzing the discrepancies with the results of Forsyth et al. [13] we can conclude that the TBFN is somewhat seductive. Allowing aggressive step sizes it appears as a fast and rather comfortable procedure. However, in spite of iteration convergence, TBFN results can possess large time truncation errors, unless the target change parameters, and accordingly the step sizes, are kept sufficiently small. The selection of these parameters is empirical. In contrast, the PCOSN results are based on temporal discretization requirements. Considering the examples analyzed in this work we can draw the following conclusions:

(1) The primary variable switching technique is able to handle any value of (negative) initial pressures. The scheme remains mass-conservative for an arbitrary time step size (see Section 8.1.2).

(2) The primary variable switching technique provides a much better convergence behavior compared to both the mixed ($\psi - s$)-form and the standard $\psi$-form of the Richards equation. This is independent of the used time marching scheme (*cf.* Table 3). The efficiency of the primary variable switching technique grows with decreasing initial pressure $\psi^0$. The acceleration usually ranges between 2 and 10, sometimes even more. The primary variable switching technique seems to be the only practical way to tackle unsaturated flow processes at very dry initial conditions.

(3) The time marching procedure and iteration control influence significantly the solution efficiency. The adaptive PCOSN scheme satisfies a predefined temporal discretization bound and usually requires more time and Newton steps at dry initial conditions than the TBFN scheme. Depending on the problem and the control parameter enforced, the TBFN can be three to six times faster than PCOSN (Sections 8.1.1, 8.1.2, 8.4.1, 8.4.2).

(4) As soon as a fully saturated zone occurs (perched water table problems) the PCOSN becomes superior and more effective (Sections 8.2 and 8.3), unless a more complex time control is used for the TBFN.

(5) The TBFN procedure does not guarantee a temporal accuracy. Resulting errors can be significant and sometimes larger than spatial discretization effects (see Figs. 16–18; Figs. 3, 4 and 9). TBFN sacrifices temporal accuracy in favor of accelerated convergence.

(6) The time marching schemes are formulated for both a first-order accurate (FE/BE) and a second-order accurate (AB/TR) strategy. For the primary variable switching technique we find that the fully implicit FE/BE scheme is more robust and should normally be preferred. This is in contrast to a standard $\psi$-form, where the higher-order AB/TR scheme works very well. In the primary variable switching technique numerical disturbances for the AB/TR scheme can be generated by the acceleration vectors $\dot{\boldsymbol{X}}$ occurring in both the Jacobian and the residual (see Eqs. (20), (A1) and (B1)). To improve the situation and gain further insights, additional investigations are required for higher-order schemes applied to the primary variable switching technique.

(7) The upstream weighting technique used in this work is easy to implement for the finite element

method. It can eliminate spurious local maxima and minima in coarse meshes (Figs. 20 and 21). Upstreaming is associated with a phase lead error which can often be tolerated with respect to the remaining errors.

(8) In simulating capillary barrier problems the situation is rather mixed. If the initial pressure is moderate there is no need to prefer variable switching since the primary interest is in steady-state solutions. Otherwise, if perched water develops, the convergence behavior is quite poor for a TBFN iteration strategy and a PCOSN method becomes more effective. On the other hand, for very dry conditions with no perched waters the variable switching technique with the TBFN strategy cannot be beaten (Sections 8.5.1 and 8.5.2).

(9) The deviatory convergence criteria in form of $L_2$ (43) and $L_\infty$ (44) error norms are basically employed in the one-step Newton (PCOSN) scheme. The same criteria are utilized for the TBFN in the present work. In the examples it has been shown that the overall iteration process can be reasonably controlled and global mass balance errors remain sufficiently small. However, in certain situations (e.g., sharp parameter contrasts) we find a stronger criterion in form of the maximum $L_\infty$ norm is to be preferred to limit the global mass balance errors below a certain level, so as done in the capillary barrier simulations. Here, the direct (or additional) use of a residual convergence criterion such as Eq. (53) would improve the global mass balance control (for sure, one would terminate the Newton iteration only if the residual satisfies the roundoff error). Such a criterion can be simply incorporated into the TBFN. But for the predictor–corrector technique, the Newton iteration can no longer be restricted to only one step and, as a result, two user-specified tolerances are necessary. This is a subject of further investigations.

The above simulations refer to 2D (1D) problems for which comparable results are available. However, the schemes discussed in this paper have been developed for both 2D and 3D applications. The present computations were performed with the FEFLOW® simulator [9].

**Acknowledgements**

The authors thank P.A. Forsyth and a second anonymous reviewer for their suggestions and critical discussion of this work. We thank J. Fuhrmann (Weierstrass Institute for Applied Analysis and Stochastics) for helpful comments. We are indebted to C. Kaiser (WASY Ltd) for his assistance with the analytical solution of capillary barrier problems.

## Appendix A. Jacobian $J^\psi$ for the pressure head $\psi$ as primary variable

The derivative of the residual (20) with respect to the pressure head $\Psi_\tau^{n+1}$ at the new time plane $n+1$ and the current iterate $\tau$ yields the following expressions $(I,J,L=1,\ldots,M)$:

$$\begin{aligned} J_{IJ}^{\psi}\left(\Psi_\tau^{n+1},\boldsymbol{s}_\tau^{n+1}\right) &= \frac{\partial R_I^{n+1}\left(\Psi_\tau^{n+1},\boldsymbol{s}_\tau^{n+1}\right)}{\partial \psi_{\tau J}^{n+1}} \\ &= J_{IJ}^{\psi 1}+J_{IJ}^{\psi 2}+J_{IJ}^{\psi 3}+J_{IJ}^{\psi 4}-J_{IJ}^{\psi 5} \\ &= \frac{\partial O_{IL}\left(\Psi_\tau^{n+1}\right)}{\partial \psi_{\tau J}^{n+1}}\left[\frac{\sigma}{\Delta t_n}\left(\psi_{\tau L}^{n+1}-\psi_L^n\right)-(\sigma-1)\dot{\psi}_L^n\right] \\ &\quad + B_{IL}\frac{\partial s_{\tau L}^{n+1}}{\partial \psi_{\tau J}^{n+1}}\frac{\sigma}{\Delta t_n}+\psi_{\tau L}^{n+1}\frac{\partial K_{IL}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\partial \psi_{\tau J}^{n+1}}+\frac{\sigma O_{IJ}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\Delta t_n} \\ &\quad + K_{IJ}\left(\Psi_\tau^{n+1}\right)-\frac{\partial F_I\left(\boldsymbol{s}_\tau^{n+1}\right)}{\partial \psi_{\tau J}^{n+1}}. \end{aligned} \tag{A1}$$

The partial Jacobians in Eq. (A1) are obtained as follows

$$J_{IJ}^{\psi 1}=\sum_e\int_{\Omega_e} N_I S_o C_{\tau J}^{n+1}\delta_{IJ}\left[\frac{\sigma}{\Delta t_n}\left(\psi_{\tau J}^{n+1}-\psi_J^n\right)-(\sigma-1)\dot{\psi}_J^n\right]$$

no summation over $I$ and $J$ (A2)

$$J_{IJ}^{\psi 2}=\sum_e\int_{\Omega_e} N_I \varepsilon C_{\tau J}^{n+1}\delta_{IJ}\frac{\sigma}{\Delta t_n}$$

no summation over $I$ and $J$ (A3)

$$J_{IJ}^{\psi 3}=\sum_e\int_{\Omega_e}\frac{\partial N_I}{\partial x_i}K_{ij}^{n+1}N_J G_{\tau J}^{n+1}\frac{\partial N_L}{\partial x_j}\psi_{\tau L}^{n+1}$$

no summation over $J$ (A4)

$$J_{IJ}^{\psi 4}=\frac{\sigma O_{IJ}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\Delta t_n}+K_{IJ}\left(\Psi_\tau^{n+1}\right) \tag{A5}$$

$$J_{IJ}^{\psi 5}=-\sum_e\int_{\Omega_e}\frac{\partial N_I}{\partial x_i}K_{ij}N_J(1+\chi)G_{\tau J}^{n+1}e_j$$

no summation over $J$ (A6)

with

$$C_{\tau J}^{n+1}=\frac{\partial s\left(\psi_{\tau J}^{n+1}\right)}{\partial \psi_{\tau J}^{n+1}} \tag{A7}$$

and

$$G_{\tau J}^{n+1}=\frac{\partial K_r\left(\psi_{\tau J}^{n+1}\right)}{\partial \psi_{\tau J}^{n+1}}. \tag{A8}$$

The derivatives $C_{\tau J}^{n+1}$ and $G_{\tau J}^{n+1}$ are given functions which can be evaluated either analytically from the parametric models (3)–(6) or numerically from cord slope approximations (Appendix C) for the known variables $s$ and $\psi$ at the iterate $\tau$, the node $J$ and the time plane $n+1$. Here, $\boldsymbol{C}_\tau^{n+1}$ is the moisture capacity

function known from the standard unsaturated modeling.

## Appendix B. Jacobian $J^S$ for the saturations as primary variable

The derivative of the residual (20) with respect to the saturation $\boldsymbol{s}_\tau^{n+1}$ at the new time plane $n+1$ and the current iterate $\tau$ yields the following expressions $(I,J,L=1,\ldots,M)$:

$$\begin{aligned} J_{IJ}^{s}\left(\boldsymbol{\Psi}_\tau^{n+1},\boldsymbol{s}_\tau^{n+1}\right) &= \frac{\partial R_I^{n+1}\left(\boldsymbol{\Psi}_\tau^{n+1},\boldsymbol{s}_\tau^{n+1}\right)}{\partial s_{\tau J}^{n+1}} \\ &= J_{IJ}^{s1}+J_{IJ}^{s2}+J_{IJ}^{s3}+J_{IJ}^{s4}-J_{IJ}^{s5} \\ &= \frac{\partial O_{IL}\left(\boldsymbol{\Psi}_\tau^{n+1}\right)}{\partial s_{\tau J}^{n+1}}\left[\frac{\sigma}{\Delta t_n}\left(\psi_{\tau L}^{n+1}-\psi_L^n\right)-(\sigma-1)\dot{\psi}_L^{\,n}\right] \\ &\quad +B_{IJ}\frac{\sigma}{\Delta t_n}+\psi_{\tau L}^{n+1}\frac{\partial K_{IL}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\partial s_{\tau J}^{n+1}} \\ &\quad +\left(\frac{\sigma O_{IL}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\Delta t_n}+K_{IL}\left(\boldsymbol{\Psi}_\tau^{n+1}\right)\right)\frac{\partial\psi_{\tau L}^{n+1}}{\partial s_{\tau J}^{n+1}}-\frac{\partial F_I\left(\boldsymbol{s}_\tau^{n+1}\right)}{\partial s_{\tau J}^{n+1}}. \end{aligned} \tag{B1}$$

The partial Jacobians in Eq. (B1) are obtained as follows

$$J_{IJ}^{s1}=\sum_e\int_{\Omega_e}N_I S_o\delta_{IJ}\left[\frac{\sigma}{\Delta t_n}\left(\psi_{\tau J}^{n+1}-\psi_J^n\right)-(\sigma-1)\dot{\psi}_J^{\,n}\right]$$

no summation over $I$ and $J$ (B2)

$$J_{IJ}^{s2}=B_{IJ}\frac{\sigma}{\Delta t_n} \tag{B3}$$

$$J_{IJ}^{s3}=\sum_e\int_{\Omega_e}\frac{\partial N_I}{\partial x_i}K_{ij}N_J G_{\tau J}^{n+1}\hat{C}_{\tau J}^{n+1}\frac{\partial N_L}{\partial x_j}\psi_{\tau L}^{n+1}$$

no summation over $J$ (B4)

$$J_{IJ}^{s4}=\left(\frac{\sigma O_{IJ}\left(\boldsymbol{s}_\tau^{n+1}\right)}{\Delta t_n}+K_{IJ}\left(\boldsymbol{\Psi}_\tau^{n+1}\right)\right)\hat{C}_{\tau J}^{n+1}$$

no summation over $J$ (B5)

$$J_{IJ}^{s5}=-\sum_e\int_{\Omega_e}\frac{\partial N_I}{\partial x_i}K_{ij}N_J(1+\chi)G_{\tau J}^{n+1}\hat{C}_{\tau J}^{n+1}e_j$$

no summation over $J$ (B6)

with the inverse moisture capacity

$$\hat{C}_{\tau J}^{n+1}=\frac{\partial\psi\left(s_{\tau J}^{n+1}\right)}{\partial s_{\tau J}^{n+1}}=\frac{1}{C_{\tau J}^{n+1}} \tag{B7}$$

which can be either derived analytically from Eqs. (3) and (5) or numerically by using cord slope approximations (Appendix C). Notice, it is necessary to use the pressure head $\psi$ instead of the hydraulic head $h$ to evaluate the moisture capacity functions $C_{\tau J}^{n+1}$ and $\hat{C}_{\tau J}^{n+1}$. Actually, $C_{\tau J}^{n+1}$ can also be expressed by $h$ since $\partial s/\partial\psi=\partial s/\partial h$, but the inverse moisture capacity $\hat{C}_{\tau J}^{n+1}$ is not simply invertible for $h$ because $\partial\psi/\partial s=\partial h/\partial s-\partial z/\partial s$.

## Appendix C. Cord slope approximations of saturation derivatives

In contrast to analytical derivatives in form of the moisture capacity $\boldsymbol{C}_\tau^{n+1}$ (A7) and its inverse $\hat{\boldsymbol{C}}_\tau^{n+1}$ (B7) cord slope approximations can be useful and effective. Within the predictor–corrector one-step Newton scheme proposed here the derivative terms are evaluated by using the predicted solutions (35), (36) for the current time plane $n+1$. For instance, a simple first-order accurate finite difference approximation of $\boldsymbol{C}_\tau^{n+1}$ would lead to

$$C_{\tau I}^{n+1}=\frac{s_{\tau I}^{n+1}-s_I^n}{\psi_{\tau I}^{n+1}-\psi_I^n}. \tag{C1}$$

Since only one iteration per time step is employed for the present predictor–corrector one-step Newton technique the iterates indicated by the subscript $\tau$ can be replaced by the predictors denoted by the subscript $p$. This yields

$$C_{pI}^{n+1}=\frac{s_{pI}^{n+1}-s_I^n}{\psi_{pI}^{n+1}-\psi_I^n}. \tag{C2}$$

It can be easily seen that this derivative is nothing more than the quotient of the acceleration vectors (35) for the saturation and the pressure head

$$C_{pI}^{n+1}=\frac{\dot{s}_I^n}{\dot{\psi}_I^{\,n}} \tag{C3}$$

which represents a cord slope approximation of the saturation derivative applied to the first-order accurate BE scheme.

A corresponding second-order accurate cord slope approximation suited for the TR scheme can be similarly derived using Eq. (41) as

$$C_{pI}^{n+1}=\frac{\Delta t_{n-1}^2\left(s_{pI}^{n+1}-s_I^n\right)+\Delta t_n^2\left(s_I^n-s_I^{n-1}\right)}{\Delta t_{n-1}^2\left(\psi_{pI}^{n+1}-\psi_I^n\right)+\Delta t_n^2\left(\psi_I^n-\psi_I^{n-1}\right)}. \tag{C4}$$

The cord slope approximations for the inverse moisture capacity $\hat{C}_{pI}^{n+1}$ yield equivalent expressions.

Note here that limitations exist for the cord slope approximations if the denominator of Eqs. (C3) and (C4) tends to zero. Practically, below an absolute minimum difference tolerance (typically we use $10^{-18}$ for the pressure head and $10^{-8}$ for the saturation) the evaluation of the derivative becomes an analytical (exact) procedure.

## References


[1] Allen MB, Murphy C. A finite element collocation method for variably saturated flows in porous media. Numer. Methods Partial Differential Equations 1985;3:229–39.

[2] Bear J, Bachmat Y. Introduction to modeling of transport phenomena in porous media, Kluwer Academic Publishers, Dordrecht, 1991.

[3] Bixler NE. An improved time integrator for finite element analysis. Communications Appl. Num. Meths. 1989;5:69–78.

[4] Celia MA, Bouloutas ET, Zarba RL. A general mass-conservative numerical solution for the unsaturated flow equation. Water Resour. Res. 1990;26(7):1483–96.

[5] Diersch H-J. Finite element modelling of recirculating density-driven saltwater intrusion processes in groundwater. Adv. Water Resour. 1988;11:25–43.

[6] Diersch, H-JG, Kolditz O. On finite-element analysis of spatio-temporal buoyancy-driven convection processes in porous media. Calibration and Reliability in Groundwater Modelling, Proceedings of the ModelCARE 96 Conference, IAHS Publ. no. 237, 1996, 407–15.

[7] Diersch, H-JG. A shock-capturing finite-element technique for unsaturated–saturated flow and transport problems, in: V.N. Burganos et al. editors. Computational Methods in Water Resources XII, vol 1, Computational Mechanics Publications, Southampton, 1998:207–14.

[8] Diersch H-JG. Treatment of free surfaces in 2D and 3D groundwater modeling. Mathematische Geologie 1998;2:17–43.

[9] Diersch H-JG. Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. User's Manual Release 4.7, May 1998, WASY Ltd., Berlin.

[10] Diersch, H-JG, Kolditz O. Coupled groundwater flow and transport: 2. Thermohaline and 3D convection systems. Adv. Water Resour. 1998;21:401–25.

[11] Engelman, MS, Strang G, Bathe K-J. The application of quasi-Newton methods in fluid mechanics. Intern. J. Num. Meths. Engng. 1981;17:707–18.

[12] Forsyth PA, Simpson RB. A two-phase, two-component model for natural convection in a porous medium. Intern. J. Num. Meths. Fluids 1991;12:655–82.

[13] Forsyth PA, Wu, YS, Pruess K. Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. Adv. Water Resour. 1995;18:25–38.

[14] Forsyth PA, Kropinski MC. Monotonicity considerations for saturated-unsaturated subsurface flow. SIAM J. Sci. Comput. 1997;18(5):1328–54.

[15] Frind EO, Verge M. Three-dimensional modeling of groundwater flow systems. Water Resour. Res. 1978;14(5):844–56.

[16] Fuhrmann J. On numerical solution methods for nonlinear parabolic problems. Notes on Numerical Fluid Mechanics, vol. 59, Vieweg, 1997:170–80.

[17] Gresho PM, Lee RL, Sani RL. On the time-dependent solution of the incompressible Navier-Stokes equations in two and three dimensions. Preprint UCRL-83282, Lawrence Livermore Lab., University of California, 1979.

[18] Haverkamp R, Vauclin M, Touma J, Wierenga PJ, Vachaud G. Comparison of numerical simulation models for one-dimensional infiltration. Soil Sci. Soc. Am. J. 1977;41:285–94.

[19] Hillel, D. Fundamentals of Soil Physics, Academic, San Diego, CA, 1980.

[20] Hills RG, Hudson DB, Porro I, Wierenga PJ. Modeling one-dimensional infiltration into very dry soils, 1, Model development and evaluation. Water Resour. Res. 1989;25(6):1259–69.

[21] Hughes TJR. A simple scheme for developing 'upwind' finite elements. Intern. J. Num. Meths. Engng. 1978;12:1359–65.

[22] Huyakorn PS, Thomas SD, Thompson BM. Techniques for making finite elements competitive in modeling flow in variably saturated media. Water Resour. Res. 1984;20:1099–115.

[23] Huyakorn PS, Springer EP, Guvanasen V, Wadsworth TD. A three-dimensional finite element model for simulating water flow with variably saturated porous media. Water Resour. Res. 1986;22:1790–808.

[24] Huyakorn PS, Panday S, Wu YS. A three-dimensional multiphase flow model for assessing NAPL contamination in porous and fractured media, 1. Formulation. J. Contaminant Hydrology 1994;16:109–30.

[25] Ju S-H, Kung K-JS, Mass types, element orders and solution schemes for the Richards equation. Computers and Geosciences 1997;23(2):175–87.

[26] Kirkland MR, Hills RG, Wierenga PJ. Algorithms for solving Richards' equation for variably saturated soils. Water Resour. Res. 1992;28(8):2049–58.

[27] Lehmann F, Ackerer Ph. Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media. Transport in Porous Media 1998;31(3):275–92.

[28] McCord JT. Application of second-type boundaries in unsaturated flow modeling. Water Resour. Res. 1991;27(12):3257–60.

[29] Miller CT, Williams GA, Kelley CT, Tocci MD. Robust solution of Richards' equation for nonuniform porous media. Water Resour. Res. 1998;34(10):2599–610.

[30] Milly PCD. A mass-conservative procedure for time-stepping in models of unsaturated flow. Adv. Water Resour. 1985;8:32–6.

[31] Neuman SP. Saturated–unsaturated seepage by finite elements. J. Hydraul. Div., ASCE 1973;99(HY12):2233–50.

[32] Nguyen H. A Petrov-Galerkin finite element scheme for one-dimensional water flow and solute transport processes in the unsaturated zone. in: A.A. Aldama et al. (Eds), Computational Methods in Water Resources XI, vol. 1, Computational Mechanics Publications, Southampton, 1996:559–66.

[33] Oldenburg CM, Pruess K. On numerical modeling of capillary barriers. Water Resour. Res. 1993;29(4):1045–56.

[34] Panday S, Forsyth PA, Falta RW, Wu Y-S, Huyakorn PS. Considerations for robust compositional simulations of subsurface nonaqueous phase liquid contamination and remediation. Water Resour. Res. 1995;31(5):1273–89.

[35] Paniconi C, Aldama AA, Wood EF. Numerical evaluation of iterative and noniterative methods for the solution of the nonlinear Richards equation. Water Resour. Res. 1991;27(6):1147–63.

[36] Paniconi C, Putti M. A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems. Water Resour. Res. 1994;30(12):3357–74.

[37] Rathfelder K, Abriola LM. Mass conservative numerical solutions of the head-based Richards equation. Water Resour. Res. 1994;30(9):2579–86.

[38] Romano N, Brunone B, Santini A. Numerical analysis of one-dimensional unsaturated flow in layered soils. Adv. Water Resour. 1998;21:315–24.

[39] Ross B. The diversion capacity of capillary barriers. Water Resour. Res. 1990;26(10):2625–29.

[40] Segol G. Classic Groundwater Simulations: Proving and Improving Numerical Models, PTR Prentice-Hall, Englewood Cliffs, NJ, 1994.

[41] Simunek J, Vogel T, Van Genuchten MTh. The SWMS-2D code for simulating water flow and solute transport in two-dimensional variably saturated media. Research Report No. 126, US Salinity Lab., Riverside, CA., 1992.

[42] Tocci MD, Kelley CT, Miller CT. Accurate and economical solution of the pressure-head form of Richards' equation by the method of lines. Adv. Water Resour. 1997;20:1–14.

[43] Van Genuchten MTh. Numerical solutions of the one-dimensional saturated-unsaturated flow equation. Research Rep. No. 78-WR-09, Water Resources Program, Princeton University, 1978.
[44] Van Genuchten MTh. A comparison of numerical solutions of the one-dimensional unsaturated–saturated flow and mass transport equations. Adv. Water Resour. 1982;5:47–55.
[45] Van der Vorst HA. Bi-CGSTAB: A fast and smoothly convergent variant of BiCG for the solution of nonsymmetric linear systems. SIAM J. Sci. Stat. Comp. 1992;13:631–44.
[46] Vogel T, Huang K, Zhang R, Van Genuchten MTh. The HYDRUS code for simulating one-dimensional water flow, solute transport, and heat movement in variably-saturated media. Research Report No. 140, U.S. Salinity Lab., Riverside, CA., 1996.
[47] Webb SW. Generalization of Ross' tilted capillary barrier diversion formula for different two-phase characteristic curves. Water Resour. Res. 1997;33(8):1855–59.