

Error propagation in the Newton-based solution control of unsaturated flow

HANS-JÖRG G. DIERSCH

*WASY Institute for Water Resources Planning and Systems Research Ltd,
Waltersdorfer Strasse 105, D-12526 Berlin, Germany
e-mail: h.diersch@wasy.de*

PIERRE PERROCHET

Centre d'Hydrogéologie, Université de Neuchâtel, Rue Emile-Argand 11, CH-2007 Neuchâtel, Switzerland

Abstract The Newton method represents the numerical core of the primary variable switching technique (PVST) which has been shown to be superior to conventional approaches in both unsaturated flow and multiphase flow modelling. In the context of PVST, empirically controlled strategies in time are rather common, where the Newton convergence is attempted for a possibly large step size. This technique is known as the target-based full Newton (TBFN) time stepping strategy. In comparison to adaptive techniques satisfying a predefined discretization error, the TBFN results can be inaccurate in spite of the convergence achieved in the Newton method. The present paper aims to analyse the cause of discrepancies in simulating unsaturated flows. This is done by comparison of analytical solutions which are based on exponential constitutive laws.

INTRODUCTION

In contrast to Picard iteration schemes common in solving the Richards' equation for unsaturated flow in porous media, the Newton method in combination with appropriate strategies can reduce the solution effort by orders. This has been shown by Forsyth *et al.* (1995) who introduced the idea of the primary variable switching technique (PVST) to saturated–unsaturated flow simulations. The major advantages of PVST are that it is: (a) unconditionally mass-conservative with respect to the time step size, (b) very effective and robust for dry initial conditions, (c) a Newton-based iteration method with quadratic convergence, and (d) a general analysis method suitable for single- and multiphase flow problems.

To control the overall iteration process Forsyth *et al.* (1995) preferred an empirical target-based full Newton (TBFN) time stepping strategy. Recently, Diersch & Perrochet (1999) compared the TBFN with an adaptive temporally error-controlled predictor-corrector technique one-step Newton scheme (PCOSN). In their extensive numerical benchmark analysis Diersch & Perrochet (1999) found that, in spite of the iteration convergence achieved, TBFN results can rather depart from PCOSN findings, unless the target change parameters, and accordingly the step sizes, are kept sufficiently small. In continuing the analysis the present paper aims at a quantification of the resulting errors using analytical solutions for the Richards' equation based on exponential saturation–pressure and conductivity–pressure relationships.

MODEL EQUATIONS

The present finite element model is based on the Richards' equation written in the following form:

$$R(s, \psi) = S_0 \cdot s(\psi) \frac{\partial \psi}{\partial t} + \varepsilon \frac{\partial s(\psi)}{\partial t} - \nabla \cdot \{K_r(\psi) \mathbf{K}[\nabla \psi + (1 + \chi)e]\} - Q = 0 \tag{1}$$

which has to be solved either for ψ or s . In equation (1),

- ψ pressure head, ($\psi > 0$ saturated medium, $\psi \leq 0$ unsaturated medium);
- $s(\psi)$ saturation, ($0 < s \leq 1$, $s = 1$ if medium is saturated);
- t time;
- S_0 specific storage due to fluid and medium compressibility;
- ε porosity;
- $K_r(\psi)$ relative hydraulic conductivity ($0 < K_r \leq 1$, $K_r = 1$ if saturated at $s = 1$);
- \mathbf{K} tensor of hydraulic conductivity for the saturated medium (anisotropy);
- χ buoyancy coefficient including fluid density effects;
- e gravitational unit vector;
- Q specific mass supply;
- R residual.

Constitutive relationships are additionally required (a) for the saturation s as a function of the pressure (capillary) head ψ , as well as its inverse, the pressure head ψ as a function of the saturation s , and (b) for the relative hydraulic conductivity K_r as a function of either the pressure head ψ or the saturation s :

$$s = f(\psi) \quad \psi = f^{-1}(s) \tag{2}$$

$$K_r = g(\psi) = g^*(s)$$

Here, van Genuchten or Brooks-Corey parametric models are common (cf. Diersch & Perrochet, 1999). Instead, if exponential constitutive laws are preferred in the form:

$$\frac{s - s_r}{1 - s_r} = s_e = \begin{cases} \exp[\alpha(\psi + \psi_a)] & \text{for } \psi < \psi_a \\ 1 & \text{for } \psi \geq \psi_a \end{cases} \tag{3}$$

$$K_r = s_e$$

analytical solutions of the nonlinear Richards' equation can be derived. In equation (3), ψ_a is the air entry pressure head, s_r is the residual saturation and $\alpha > 0$ is a constant.

NEWTON METHOD AND PVST

The discretized form of the basic Richards' equations (1) yields:

$$\mathbf{R}^{n+1}(\mathbf{X}) = 0 \tag{4}$$

to be solved for a primary variable:

$$\mathbf{X} \in (\psi, s) \tag{5}$$

which can be either ψ or s at the new time level $n + 1$. Applying the Newton method to

equation (4) we solve:

$$\mathbf{J}^X(\boldsymbol{\psi}_\tau^{n+1}, \mathbf{s}_\tau^{n+1}) \Delta \mathbf{X}_\tau^{n+1} = -\mathbf{R}_\tau^{n+1}(\boldsymbol{\psi}, \mathbf{s}) \quad (6)$$

with the increment:

$$\Delta \mathbf{X}_\tau^{n+1} = \mathbf{X}_{\tau+1}^{n+1} - \mathbf{X}_\tau^{n+1} \quad (7)$$

and the Jacobian \mathbf{J}^X expressed in indicial notation as:

$$J_{IJ}^X(\boldsymbol{\psi}_\tau^{n+1}, \mathbf{s}_\tau^{n+1}) = \frac{\partial R_I^{n+1}(\boldsymbol{\psi}_\tau^{n+1}, \mathbf{s}_\tau^{n+1})}{\partial X_{\tau J}^{n+1}} \quad (8)$$

where τ denotes the iteration number. The PVST selects the primary variable in a dynamic manner depending on inner nodal criteria of the solutions, $\boldsymbol{\psi}$ or \mathbf{s} . The derivatives of the Jacobian can be easily switched between $\boldsymbol{\psi}$ and \mathbf{s} in accordance with the computational requirements. Their computations can be done either analytically or numerically.

THE NITTY-GRITTY

Generally, the control of the solution of the resulting highly nonlinear matrix system (equation (6)) is a tricky matter. Both the choice of the time step size Δt_n and the iteration control of the Newton scheme significantly influence the success and the efficiency of the simulation. In the PCOSN scheme (Diersch & Perrochet, 1999) the nonlinear matrix system is linearized by the predictor solutions. Temporal truncation errors can be easily estimated by evaluating predictor and corrector solutions which are the basis of an adaptive error-controlled time stepping and iteration strategy. In contrast, the TBFN (Forsyth *et al.*, 1995) does not consider temporal truncation errors in the time and iteration control. 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.

An important aspect of the iterative solution via the PCOSN and TBFN schemes is the choice of an appropriate convergence criterion. Limiting the temporal discretization errors deviatory (change) error measures $\|\mathbf{d}_\tau^{n+1}\|_{L_p}$ are the controlling criteria, which are functions of the solution differences $\mathbf{d}_\tau^{n+1} \sim \Delta \mathbf{X}_\tau^{n+1}$:

$$\|\mathbf{d}_\tau^{n+1}\|_{L_p} < \delta \quad (9)$$

where δ is a user-specified deviatory error tolerance. Here, weighted RMS L_2 and maximum L error norms can be chosen. Commonly, in the Newton method the deviatory error criterion (equation (9)) represents a standard test to terminate the iteration within the time step. In the PCOSN the temporal truncation and the Newton termination error measures are equivalently used. As a result, only one error criterion and one Newton step per time step become necessary ($\tau = 1$). As an alternative to the deviatory error estimate $\|\mathbf{d}_\tau^{n+1}\|_{L_p}$, the residual $\|\mathbf{R}_\tau^{n+1}\|_{L_p}$ may be directly controlled, such as:

$$\|\mathbf{R}_\tau^{n+1}\|_{L_p} < \delta_2 \|\mathbf{F}^{n+1}\|_{L_p} \tag{10}$$

where an additional error tolerance δ_2 appears and an appropriate normalization of the residual (here with respect to the external supply \mathbf{F}^{n+1}) is required. In the TBFN deviator errors $\|\mathbf{d}_\tau^{n+1}\|_{L_p}$ and residual errors $\|\mathbf{R}_\tau^{n+1}\|_{L_p}$ can be alternatively employed.

Instead of a one-step Newton control as done in the PCOSN the predictor-corrector technique can also be extended to a multiple step Newton (PCMSN) strategy satisfying both criteria (9) and (10). To measure the global balance error we introduce:

$$\mathfrak{R}(T) = \frac{\int_{t=0}^T \|\mathbf{R}_\tau(t)\|_{L_2} dt}{\int_{t=0}^T \|\mathbf{F}(t)\|_{L_2} dt} \tag{11}$$

for the ‘‘accumulated loss’’ of mass with respect to the total external supply over the entire simulation period $(0, T)$.

ANALYTICAL SOLUTION

In one dimension the Richards’ equation (1):

$$\varepsilon \frac{\partial s(\psi)}{\partial t} - \frac{\partial}{\partial z} \left[K_r(\psi) K \left(\frac{\partial \psi}{\partial z} - 1 \right) \right] = 0 \tag{12}$$

can be transformed into the linear advective–dispersive equation of the form:

$$\frac{\partial s}{\partial t} + \frac{K}{\varepsilon(1-s_r)} \frac{\partial s}{\partial z} - \frac{K}{\varepsilon\alpha(1-s_r)} \frac{\partial^2 s}{\partial z^2} = 0 \tag{13}$$

for the exponential constitutive law (3) by using the following assumptions: $S_0 = \psi_a = Q = \chi = 0$, where z is oriented downward in the direction of gravity.

With $s(z,0) = s_i$ and $s(0,t) = s_0$ the analytical solution is:

$$s(z,t) = s_i + \frac{(s_0 - s_i)}{2} \left\{ \operatorname{erfc} \left(\frac{z - \frac{Kt}{\varepsilon(1-s_r)}}{2 \sqrt{\frac{Kt}{\varepsilon\alpha(1-s_r)}}} \right) + e^{\alpha z} \operatorname{erfc} \left(\frac{z + \frac{Kt}{\varepsilon(1-s_r)}}{2 \sqrt{\frac{Kt}{\varepsilon\alpha(1-s_r)}}} \right) \right\} \tag{14}$$

It can be easily seen from equation (13) that with large α the problem is dominated by advection. Otherwise, considering a fully implicit time discretization the temporal numerical dispersion can be estimated as:

$$D_{\text{numdisp}}^{n+1} \approx \frac{\Delta t_n}{2} \left[\frac{K}{\varepsilon(1-s_r)} \right]^2 \tag{15}$$

TEST CASE

The problem is described in Fig. 1. For the lower boundary a free drainage-type boundary condition is applied (Diersch, 1998). The 6 m column is discretized by 120 linear finite elements, so the nodal spacing becomes $\Delta z = 5$ cm.

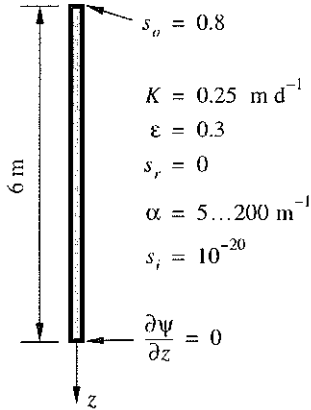


Fig. 1 Sketch of the test problem.

Newton control by the deviatory error criterion (equation (9))

The computed saturation profiles for two α -parameters in comparison with the analytical solution are shown in Fig. 2. Large conservation errors are observed for the TBFN if the number of time and accordingly Newton steps become small. For $\alpha = 5 \text{ m}^{-1}$, 20 and 61 time steps (49 and 116 Newton steps) are needed for the constraints $\Delta t_{max} = 0.2$ day and $\Delta t_{max} = 0.05$ day, respectively. It leads to the total integral balance errors (equation (11)) of \mathcal{R} (3 day) $\approx 160\%$ and $\approx 140\%$, respectively. In contrast, the

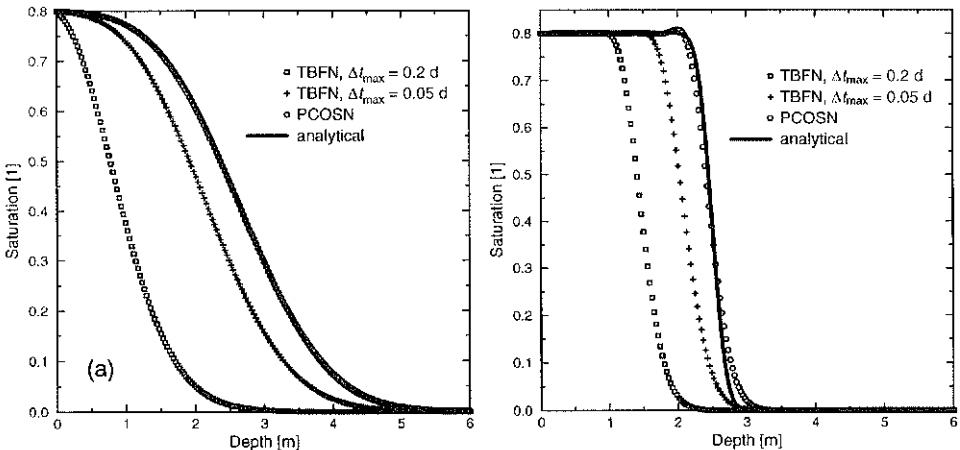


Fig. 2 Saturation profiles at $t = 3$ day for (a) $\alpha = 5 \text{ m}^{-1}$, and (b) $\alpha = 200 \text{ m}^{-1}, \delta = 10^{-4}$, with the deviatory error criterion (9) for the Newton control; aggressive target change parameters are used for the TBFN with a maximum time step constraint Δt_{max} .

PCOSN took 240 variable time and Newton steps resulting an acceptable balance error of only \mathfrak{R} (3 day) $\approx 0.06\%$ with a very good agreement with the analytical solution. Similar results appear for $\alpha = 200 \text{ m}^{-1}$, where the TBFN gives \mathfrak{R} (3 day) $\approx 7\%$ (74 time and 344 Newton steps) and \mathfrak{R} (3 day) $\approx 2\%$ (107 time and 301 Newton steps), respectively, while the PCOSN obtains \mathfrak{R} (3 day) $\approx 0.09\%$ after 360 time and Newton steps.

Newton control by the residual error criterion (equation(10))

As outlined in Fig. 3 the conservative problems disappears for the TBFN if the residual error criterion (equation (10)) is used with $\delta_2 = 10^{-4}$. The adaptive PCMSN and the TBFN give comparable results which agree quite well with the analytical solution. Here, the PCMSN is now controlled by two the criteria (9) and (10): $\delta = 10^{-4}$ for the time adaptation and $\delta_2 = 10^{-4}$ for the Newton termination, where more than one Newton step per time increment can occur. The TBFN needed 49 time (279 Newton) and 133 time (502 Newton) steps for $\alpha = 5 \text{ m}^{-1}$ and $\alpha = 200 \text{ m}^{-1}$, respectively, achieving total balance errors of \mathfrak{R} (3 day) $\approx 0.0006\%$ and \mathfrak{R} (3 day) $\approx 0.001\%$, respectively. As expected, the PCMSN took more time steps. We found 164 time (268 Newton) steps with \mathfrak{R} (3 day) $\approx 0.01\%$ and 165 time (362 Newton) with \mathfrak{R} (3 day) $\approx 0.001\%$ for $\alpha = 5 \text{ m}^{-1}$ and $\alpha = 200 \text{ m}^{-1}$, respectively.

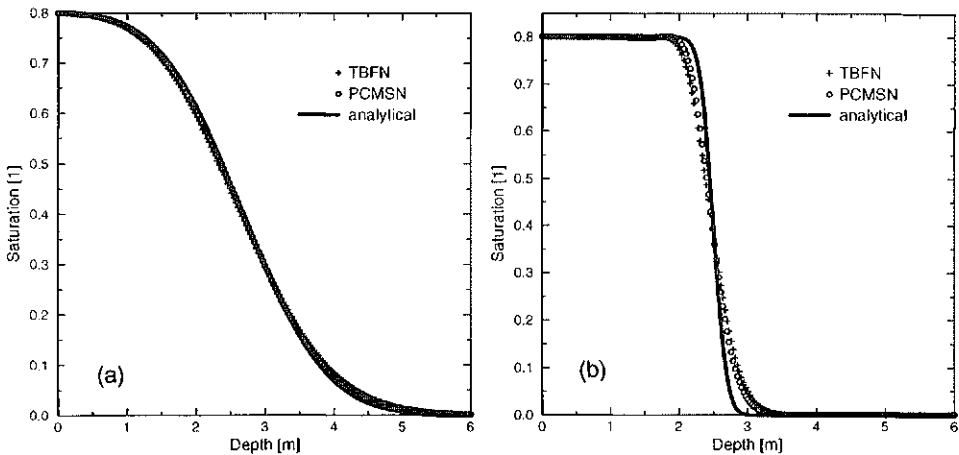


Fig. 3 Saturation profiles at $t = 3$ day for (a) $\alpha = 5 \text{ m}^{-1}$, and (b) $\alpha = 200 \text{ m}^{-1}$, $\delta = \delta_2 = 10^{-4}$ with the residual error criterion (10) for the Newton control.

CONCLUSIONS

For the TBFN the residual error criterion should be preferred over standard deviation tests to avoid the conservation errors, as long as the target change parameters allow large steps. On the other hand, the adaptive PCOSN scheme sufficiently controls the solution process by limiting time truncation errors and an additional residual test, as done in the PCMSN scheme, is not necessary in most cases.

REFERENCES

- Diersch, H.-J. G. (1998) Treatment of free surfaces in 2D and 3D groundwater modeling. *Math. Geologie* **2**, 17–43.
- Diersch, H.-J. G. & Perrochet, P. (1999) On the primary variable switching technique for simulating unsaturated-saturated flows. *Adv. Wat. Resour.* **23**(3), 271–301.
- Forsyth, P. A., Wu, Y. S. & Pruess, K. (1995) Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. *Adv. Wat. Resour.* **18**(1), 25–38.