Noniterative time stepping schemes with adaptive truncation error control for the solution of Richards equation

Dmitri Kavetski, Philip Binning, and Scott W. Sloan
School of Engineering, University of Newcastle, Callaghan, New South Wales, Australia

Received 15 June 2001; revised 7 June 2002; accepted 7 June 2002; published 24 October 2002.

[1] Noniterative implicit time stepping schemes with adaptive temporal truncation error control are developed for the solution of the pressure form of Richards equation. First- and second-order linearizations of an adaptive backward Euler/Thomas-Gladwell formulation are introduced and are shown to constrain the temporal truncation errors near a user-prescribed tolerance and maintain adequate mass balance. Numerical experiments demonstrate that accurate noniterative linearizations achieve cost-effective solutions of problems where soils are described by highly nonlinear and nonsmooth constitutive functions. For these problems many conventional iterative solvers fail to converge. The noniterative formulations are considerably more efficient than analogous time stepping schemes with iterative solvers. The second-order noniterative scheme is found to be more efficient than the first-order noniterative scheme. The proposed adaptive noniterative algorithms can be easily incorporated into existing backward Euler software, which is widely used for Richards equation and other nonlinear PDEs. INDEX TERMS: 1866 Hydrology: Soil moisture; 1875 Hydrology: Unsaturated zone; 3210 Mathematical Geophysics: Modeling; 3230 Mathematical Geophysics: Numerical solutions; KEYWORDS: Richards equation, noniterative time stepping, adaptive integration, Thomas-Gladwell scheme


1. Introduction

[2] Flows through variably saturated porous media are typically described by Richards equation, a highly nonlinear parabolic partial differential equation (PDE) that can be cast in several forms [Huyakorn and Pinder, 1985]. The pressure form of Richards equation is given by

\[
\frac{C(y) + \theta(y)S_y}{\phi} \frac{\partial y}{\partial t} - \nabla \cdot K(y) \nabla y + \frac{\partial K(y)}{\partial z} = 0, \tag{1}
\]

where \(y\) is the pressure head [L], \(\theta(y)\) is the volumetric moisture content [], \(t\) is time [T], \(z\) is the (positive downward) depth [L], \(\phi\) is the porosity [], \(S_y\) is the specific storativity [L⁻¹], \(K(y)\) is the hydraulic conductivity [L/T], and \(C(y) \equiv \partial \theta/\partial y\) is the specific moisture capacity [L⁻¹].

[3] Numerical methods for Richards equation have attracted considerable research attention and are widely used in practical simulations of subsurface processes. However, recent studies show that standard numerical schemes cannot satisfactorily solve certain flow problems, e.g., the saturation of initially dry soils with nonuniform pore size distribution [Miller et al., 1998]. This study explores the advantages of noniterative adaptive time stepping schemes for Richards equation and presents a simple yet effective method that accurately and cost-effectively approximates these difficult problems and also improves the computational efficiency in other cases. The proposed scheme is closely related to backward Euler methods and hence can be used to improve existing software for practical subsurface simulations. Indeed, “even the old workhorse is more nimble with new horseshoes” [Press et al., 1992, p. 706].

[4] Until recently, numerical methods for Richards equation were mainly limited to simple time stepping schemes coupled with finite difference or finite element spatial approximations [Huyakorn and Pinder, 1985]. The time approximations included backward Euler and related schemes [e.g., Celia et al., 1990; Paniconi et al., 1991].

[5] A fundamental advance in the numerical analysis of Richards equation was the introduction of adaptive algorithms, which adjust to the behavior of the solution and are generally more reliable and efficient than uncontrolled approaches. Adaptive spatial approximations for Richards equation include a hierarchic finite element scheme [Abriola and Lang, 1990] and a front-tracking scheme [Griffoli and Cohen, 1999]. Adaptive time integration methods included variable-order variable-step DASPK integrators [Tocci et al., 1997; Miller et al., 1998; Williams and Miller, 1999] and lower-order adaptive backward Euler and related schemes [Kavetski et al., 2001, 2002]. In all cases, formal truncation error control leads to considerable gains in accuracy and efficiency over fixed step and heuristic time stepping algorithms and also improves the mass balance of schemes based on equation (1).

[6] Unconditional stability is an important property of an effective time stepping scheme for Richards equation, due to the stiffness of spatially discrete parabolic PDEs [Huyakorn and Pinder, 1985; Wood, 1990]. Indeed, most production and research codes employ the implicit Euler algorithm, which is first-order accurate but very stable.
Common iterative solvers for the implicit nonlinear systems include the Picard and Newton schemes [Paniconi et al., 1991; Paniconi and Putti, 1994]. Other Newton-type methods have also been used, including the initial-slope Newton scheme, Newton-Krylov methods, and combined Picard-Newton schemes [Fassino and Manzini, 1998; Bergamaschi and Putti, 1999; Jones and Woodward, 2000]. In practice, Picard iteration is prevalent due to its simplicity and generally acceptable performance [Lehmann and Ackerer, 1998]. However, the nonsmoothness of constitutive functions describing some soils, e.g., certain unconsolidated loams and clay loams, causes poor convergence or complete failure of Picard and Newton solvers in uncontrolled time stepping schemes. More sophisticated variable-order variable-step schemes with chord iteration solvers, as well as Newton solvers augmented with global line searches, can be used to improve convergence in these difficult simulations [Tocci et al., 1997; Miller et al., 1998]. However, unless frequent Jacobian evaluation is avoided (e.g., by maintaining a constant time step size whenever possible), iterative solvers become computationally expensive, since multiple iterations including the recalculation and inversion of the Jacobian are necessary at each time step.

Noniterative methods require a single matrix formation and inversion per time step and thus offer potential efficiency advantages over iterative schemes. For example, the noniterative implicit factored approximation has a comparable or higher efficiency than a Crank-Nicolson scheme with a Newton solver [Paniconi et al., 1991]. However, the implicit factored scheme is not easily applied to Richards equation and has difficulties at the saturated-unsaturated interface. In simpler algorithms examined by Paniconi et al. [1991], noniterative linearizations limit the temporal accuracy to first order. Despite these difficulties, noniterative integration schemes are an attractive alternative to traditional iterative methods for Richards equation and other nonlinear PDEs.

While recent work has highlighted the benefit of adaptive time stepping for the solution of Richards equation, previous studies of noniterative solvers were limited to fixed step size algorithms. Therefore this study focuses on designing a noniterative scheme embedded in an adaptive time stepping algorithm. The key merits of the proposed noniterative adaptive scheme are (1) simplicity and efficiency: a single linear solution per time step is necessary, Jacobians are not required; (2) adaptive time step variation with explicit truncation error control, second-order temporal accuracy; (3) ability to cost-effectively solve difficult flow cases with nonsmooth \( \theta(\psi) \) and \( K(\psi) \); and (4) ease of incorporation into widely used backward Euler codes.

The combination of these properties is uncommon for noniterative schemes previously applied to Richards equation. For example, Lees’ scheme, a three-level noniterative second-order scheme, suffers stability problems when solving Richards equation [Paniconi et al., 1991] and is not easily implemented in adaptive step size applications. Many single-level noniterative schemes, e.g., linearized Crank-Nicolson formulations, suffer a degradation of accuracy that can be remedied only by using high-order Jacobians [Paniconi et al., 1991].

The new scheme is benchmarked against several iterative and noniterative algorithms, including the popular mass conservative modified Picard [Celia et al., 1990] and DASPK-KAM [Miller et al., 1998] algorithms. The test problems comprise intermittently saturated flows in poorly sorted porous media. These are challenging tests for the numerical schemes, which must cost-effectively achieve adequate accuracy despite highly nonlinear and nonsmooth constitutive functions and time varying boundary conditions. The empirical assessment confirms the robustness and efficiency of the new algorithm and demonstrates the advantages of noniterative methods. In addition, adaptive time stepping based on truncation error control is shown to improve both iterative and noniterative formulations.

2. Numerical Formulation

2.1. Spatial Approximation

Finite difference and finite element spatial discretizations of nonlinear PDEs such as equation (1) lead to a system of nonlinear ordinary differential equations (ODEs), or more formally, differential-algebraic equations (DAEs),

\[ C(\psi) \frac{d\psi}{dt} + K(\psi)\psi = F(\psi), \]  

where \( \psi \) are the pressure values at the spatial nodes, \( C \) is the mass matrix, \( K \) is the conductivity matrix, and the vector \( F \) contains the gravity drainage term and boundary conditions. The structure of \( C, K, \) and \( F \) depends on the dimensionality of the problem and on the approximation method used. The case study in this paper simulates one-dimensional vertical flows with a lumped linear Galerkin finite element spatial approximation [Celia et al., 1990; Kavetski et al., 2001]. The elemental conductivity is evaluated as the arithmetic mean of the local nodal values, i.e., \( K^{(e)} = (K_1 + K_2)/2 \). Alternative methods based on geometric, harmonic, and integral means could also be used [Miller et al., 1998] but are tangential to this study. The noniterative time stepping scheme introduced in this paper is applicable to any spatial approximation scheme that generates a finite dimensional ODE/DAE system of the form (2).

2.2. Adaptive Temporal Integration

The backward Euler scheme is a first-order time stepping scheme widely used for Richards equation and other PDEs due to its simplicity, unconditional stability, and robustness. Its accuracy and efficiency can be enhanced by the following adaptive error control scheme [Sloan and Abbo, 1999; Kavetski et al., 2001, 2002]:

1. Solve the backward Euler nonlinear system

\[ C(\psi^{n+1}_{(1)}) + \Delta tK(\psi^{n+1}_{(1)})\psi^{n+1} = -K(\psi^{n+1}_{(1)})\psi^{n} + F(\psi^{n+1}_{(1)}) \]  

\[ \psi^{n+1}_{(1)} = \psi^{n} + \Delta t\psi^{n+1}, \]  

where \( \psi^{n+1}_{(1)} \) denotes the first-order backward Euler approximation.

2. Estimate the local temporal truncation errors using the derivative estimates

\[ e^{n+1} = \frac{1}{2}\Delta t(\psi^{n+1} - \psi^{n}). \]
3. Assess the local truncation errors using a mixed absolute-relative error test. A timestep is accepted if
max(\[\epsilon_{(t+1)}^{n+1} - \tau_c | y_{(t+1)}^{n+1} - \tau_s | y_{(t)}^{n+1}\) < 0 where \(\tau_c\) and \(\tau_s\)
are user prescribed absolute and relative error tolerances. The solution is then updated according to
\[
\psi_{(t)}^{n+1} = \psi^n + \frac{1}{2} \Delta t (\psi^n + \psi_{(t)}^{n+1}),
\]
where \(\psi_{(t)}^{n+1}\) is a second-order Thomas-Gladwell approximation [Thomas and Gladwell, 1988; Kavetski et al., 2002].
The output solution of the adaptive scheme is hence \(O(\Delta t^2)\)
accurate. Note that \(\epsilon_{(t+1)}^{n+1} = \psi_{(t)}^{n+1} - \psi_{(t)}^{n+1}\). If the time step is
accepted, the step size for the next time step can be computed. If the truncation error test is not satisfied, the
time step is reattempted with a reduced step size.

[14] Additional algorithm details, including initialization, are given by Kavetski et al. [2001].

[15] This study focuses on highly nonlinear nonsmooth
problems, where iterative solvers may not converge when
solving equation (3). For these solvers, the maximum number of iterations per time step is limited to \(N_{max} = 50\). If \(N_{max}\) is exceeded, the step size is reduced by a factor \(F_r = 0.25\). While different choices of \(N_{max}\) and \(F_r\) do affect
solver performance, they do not alter the qualitative conclusions of this work.

[16] This adaptive method performs well in elastoplastic
consolidation analysis [Sloan and Abbo, 1999], and its
application with iterative solvers to the moisture and mixed
forms of Richards equation was also successful [Kavetski et al., 2001, 2002]. The marginal arithmetic and memory cost
of the adaptive mechanism is negligible, especially compared
to the solution of the nonlinear system (3).

2.3. Solution of the Nonlinear Time Stepping
Equations

[17] Since the adaptive scheme is implicit, its main computational
cost is the solution of the nonlinear system (3) at
each time step. When the Picard solver is used, the implicit
system is linearized as
\[
C(\psi^{n+1,0}) + \Delta t K(\psi^{n+1,0})_| | \psi^{n+1,0} = -K(\psi^{n+1,0}) + F(\psi^{n+1,0})
\]
where the initial estimates \(\psi^{n+1,0}\) (or, equivalently, \(\psi^{n+1,0}\))
are necessary to start the iteration. Although simply reusing \(\psi^n\) (which is equivalent to \(\psi^{n+1,0} = 0\)) is predominant
[Paniconi et al., 1991, and others], it is more efficient to
generate initial estimates by extrapolation [Cooley, 1983; Huyakorn et al., 1984]. In the context of the adaptive time
stepping scheme, where time derivatives are already available, it is natural to initialize via \(\psi^{n+1,0} = \psi^n\), i.e.,
\[
\psi^{n+1,0} = \psi^n + \Delta t \psi^n.
\]
Taylor series analysis shows that the estimate (9) is \(O(\Delta t^2)\)
accurate and generally more efficient than \(\psi^{n+1,0} = \psi^n\).

2.4. Noniterative Linearization of the Implicit System

[18] Linearized (noniterative) variants of Picard and Newton
methods can be derived from a single iteration of these
schemes with \(\psi^n\) as the initial estimate [Paniconi et al.,
1991]. For example, the noniterative “modified implicit Euler” scheme [Culham and Varga, 1971] can be formulated as
\[
C(\psi^n) + \Delta t K(\psi^n) \psi^{n+1} = -K(\psi^n) \psi^n + F(\psi^n)
\]

When expressed in this form, the modified implicit Euler scheme can be supplemented with the adaptive truncation error control described above. However, Paniconi et al. [1991] showed that when applied to \(O(\Delta t^2)\) formulations, e.g., the Crank-Nicolson scheme, linearizations based on \(\psi^n\) reduce the accuracy to \(O(\Delta t)\). The \(O(\Delta t^2)\) accuracy can then be restored only with high-order Jacobians.

[19] In order to maintain the order of accuracy of the time
discretization and keep the algorithm simple, we design a
second-order linearization of the nonlinear coefficients that
does not require Jacobians. Inspired by the extrapolation
(9), the time stepping equations (3) can be linearized as follows:
\[
C(\psi^{n+1}) + \Delta t K(\psi^{n+1}) | \psi^{n+1} = -K(\psi^{n+1}) \psi^n + F(\psi^{n+1})
\]

This approach is equivalent to performing a single Picard iteration using equation (9) as the initial estimate. In order to
examine the additional linearization error introduced by equations (12) and (13), the update (6) is rearranged as a
truncated Taylor expansion about \(\psi^t\) and combined with equation (13), showing that
\[
\psi^{n+1} = \psi^{n+1} - \frac{1}{2} \Delta t \psi^n + \Delta t \psi^n + O(\Delta t^3) = \psi^{n+1} + O(\Delta t^2).
\]

When the nonlinearities in the matrices \(C\), \(K\), and \(F\) are
evaluated with \(\psi^{n+1}\), additional \(O(\Delta t^2)\) linearization error is
introduced. For example, the approximation \(C(\psi^{n+1})\) of the
specific capacity \(C(\psi)\) is second-order accurate.
\[
C(\psi^{n+1}) = C(\psi^{n+1} - \frac{1}{2} \Delta t \psi^n)
\]

Similarly, the conductivity is also approximated with \(O(\Delta t^2)\) errors, maintaining the \(O(\Delta t^2)\) accuracy of the
adaptive time stepping scheme. The second-order convergence
of the noniterative scheme is verified empirically with both uniform and variable step sizes.

[20] The stability of the scheme is also examined. When
used with linear coefficients, the scheme is equivalent to an
unconditionally stable Thomas-Gladwell scheme [Thomas
and Gladwell, 1988]. When the differential equation is
nonlinear, the noniterative scheme is expected to be stable,
since semi-implicit schemes are typically stable for
nonlinear problems [Press et al., 1992]. Numerical experimen-
tation with a range of boundary and initial conditions showed no sign of instability.

[21] It is also noted that linearizations similar to (12) and (13) can be applied in other implicit methods. The key to this noniterative strategy is to use time derivatives to approximate the nonlinearities to the same order of accuracy as the time stepping scheme and then proceed with a single correction of the selected nonlinear solver. Similar predictors, albeit within iterative solvers, have also been used in DAE schemes such as DASPK.

2.5. Noniterative Linearization and Adaptive Truncation Error Control

[22] The noniterative linearization is compatible with the adaptive error control, since the term $\Delta t^2 \dddot{\psi}/2$ in the linearization error of the coefficients $C$, $K$, and $F$ also represents the temporal error of the backward Euler scheme. Indeed, the truncation error estimate (5) can be also expressed in terms of an approximation to the second derivative

$$\varepsilon^{n+1} = \frac{1}{2} \Delta t^2 \dddot{\psi} + O(\Delta t^3),$$

(16)

since

$$\dot{\psi} = \frac{\psi^{n+1} - \psi^n}{\Delta t} = \frac{d^2 \psi}{d t^2} + O(\Delta t).$$

Expression (16) represents the leading term of the local truncation error of the backward Euler scheme, integrated over the time step. Since the magnitude of $\Delta t^2 \dddot{\psi}/2$ in equations (15) and (16) is constrained by the user’s error tolerance, adaptive time stepping will also indirectly control the accuracy of the noniterative solver.

2.6. Noniterative Algorithms and Conservation of Mass

[23] The noniterative scheme is applied directly to the pressure form of Richards equation (1), since it is difficult to satisfactorily linearize intrinsically mass conservative algorithms. Rathfelder and Abriola [1994] have shown that mass conservation in numerical approximations of equation (1) requires chord-slope approximations to the specific capacity $d\theta/d\psi$, e.g.,

$$\frac{d\theta}{d\psi} \approx \frac{\theta^{n+1} - \theta^n}{\psi^{n+1} - \psi^n}.$$  

(17)

Mass conservative approximations based on the alternative mixed form of Richards equation are equivalent to pressure-based schemes that employ equation (17) or similar chord-slope schemes [Rathfelder and Abriola, 1994]. Approximations such as (17) introduce $O(\Delta t)$ errors, and more complex higher-order variants must be used to achieve $O(\Delta t^3)$ accuracy [Kavetski et al., 2001]. In order to obtain a noniterative formulation that uses equation (17), it is necessary to linearize both $\theta^{n+1}$ and $\psi^{n+1}$. However, it then becomes difficult to meet condition (17) after a single iteration. It appears that the only way to ensure mass balance regardless of the time step size is to resort to iteration to satisfy equation (17).

[24] However, these difficulties are not critical, since temporal truncation error control maintains adequate mass balance in time integrators based on the pressure form of Richards equation [Tocci et al., 1997]. In general, explicit error control is recommended regardless of whether the scheme is conservative, and, moreover, it will be shown that noniterative schemes succeed in solving problems that cannot be satisfactorily handled using iterative mass conservative methods with standard Newton and Picard linearizations.

2.7. Incorporation Into Existing Software

[25] The proposed noniterative scheme can easily be incorporated into existing and new codes for Richards equation based on backward Euler approximations. Although most backward Euler codes compute $\psi_{n+1}^{(1)}$ directly, the derivatives $\dot{\psi}$ can be back-calculated as

$$\dot{\psi} = \frac{(\psi_{n+1}^{(1)} - \psi_n^{(1)})}{\Delta t}.$$  

(18)

The evaluation of formula (18), (13), and (6) is very cheap, especially in comparison with matrix formations and inversions. The additional $O(N_s)$ storage for $\dot{\psi}$ is also minor. The Picard solver is set to perform a single iteration per time step. If the code implements a different solver (e.g., the Newton scheme), the same procedure is followed: Evaluate equation (13) and perform a single nonlinear iteration. The second-order accuracy of the predictor (13) will be preserved by the solver correction, provided the update (6) is used for the output solution.

[26] Implicit Euler schemes are widely used in hydrologic and engineering codes for Richards equation, as well as for other nonlinear PDEs. The practical utility of the proposed techniques is therefore in the opportunity to upgrade existing software and improve practical computational standards.

3. Results and Discussion

[27] Two test problems will be considered, both simulating vertical infiltration in a 200-cm vertical column of soil described by the van Genuchten constitutive relationships

$$0(\psi) = \begin{cases} \theta_r - \theta_s, & \text{if } \psi < 0 \\ \theta_s, & \text{if } \psi \geq 0 \end{cases}$$  

(19)

$$K(\psi) = \begin{cases} K_s \left( \frac{1 - (1 - \alpha(\psi))^{n_v-1}(1 + \alpha(\psi))^{n_v}}{1 + \alpha(\psi)^{n_v/2}} \right)^2, & \text{if } \psi < 0, \\ K_s, & \text{if } \psi \geq 0 \end{cases}$$  

(20)

with $\theta_r = 0.095$, $\theta_s = 0.410$, $K_s = 6.2$ cm/d, $\alpha = 0.019$ cm$^{-1}$, $n_v = 1.31$, and $\mu_v = 1 - 1/n_v = 0.237$. The soil compressibility parameter $S_{cr/\psi}$ in equation (1) is $10^{-6}$ cm$^{-1}$. These properties correspond to unconsolidated clay loam with a nonuniform grain size distribution and were previously used by Miller et al. [1998].

[28] The material properties used in this study are challenging for numerical solvers. In general, solver conver-
gence requires a smooth continuous Jacobian of the non-linear system [Ortega and Rheinboldt, 1970]. While the van Genuchten functions are continuous for the entire range of pressure heads, their smoothness and differentiability at \( y = 0 \) depend on \( n_v \), e.g., when \( 1 < n_v/C_2 \), \( d^2\psi/dy^2 \) is undefined at \( y = 0 \). Simulation of such saturated soils using traditional fixed step Picard methods is very difficult [Miller et al., 1998].

3.1. Test Problem 1

[29] The first test problem comprises nontrivial time-dependent boundary conditions. The initial pressure profile is specified as

\[
\psi(z, t = 0) = \begin{cases} 
-5 + \frac{5.625 - 200}{0.625} - z & 0 < z \leq 0.625 \\
-200 & 0.625 < z \leq 200 
\end{cases}
\]

The lower end of the soil column is permanently saturated, \( \psi(z = 200, t) = 0 \), while a time-dependent Dirichlet condition is imposed at the top boundary.

\[
\psi(z, t) = \begin{cases} 
-5 + \frac{3 \sin(2\pi t/100000)}{10} & 0 < t \leq 100000 \\
10 & 100000 < t \leq 180000 \\
-5 + 295245 \exp(-t/18204.8) & 180000 < t \leq 300000 
\end{cases}
\]

The pressure variation described by equation (22) could represent a tidal or irrigation pattern. The pressure follows a smooth sinusoid for one full period, and then a sharp pulse saturates the upper end of the domain for 80,000 s, followed by a gradual return to unsaturated levels toward the end of the simulation. The solution is shown in Figure 1 and represents a sharp pressure front propagating through the domain. Robust nonlinear solver performance and dynamic time step size adjustment are required to handle the saturated-unsaturated interface and the changes in forcing conditions. We report the performance of selected numerical schemes for both the entire simulation and the three boundary condition periods separately.

[30] The new \( O(\Delta t^2) \) noniterative adaptive scheme is compared to three alternative numerical formulations:

1. The first is the modified Picard scheme [Celia et al., 1990]. This \( O(\Delta t) \) iterative algorithm is common in research and production codes because it guarantees conservation of mass. However, fixed step implementations of this Picard method are uneconomical for problems with nonsmooth constitutive functions [Miller et al., 1998]. In this study, the modified Picard scheme is supplemented with the adaptive error control described above. Since the modified Picard approximation is equivalent to an algorithm based on the \( O(\Delta t) \) chord-slope scheme (17), the update (6) cannot achieve \( O(\Delta t^2) \) accuracy and is omitted. Although this theoretically affects the accuracy of the error estimator (5), in practice we found the error control to be useful and far more effective than heuristic step size variation methods. The modified Picard scheme is the only intrinsically mass conservative scheme used in this study.

2. The second is a standard Picard scheme, comprising an adaptive backward Euler scheme with an iterative Picard solver. When applied to the moisture form of Richards equation, this \( O(\Delta t^2) \) algorithm is more reliable and efficient than fixed step and heuristic backward Euler and Thomas-Gladwell methods [Kavetski et al., 2002].
3. The third is a modified implicit Euler scheme (10) with adaptive error control. We consider this noniterative \( O(\Delta t) \) algorithm to further demonstrate (1) generic advantages of noniterative formulations and (2) the importance of the order of accuracy of time stepping schemes and nonlinear solvers. Since linearization (10) is \( O(\Delta t) \), the update (6) cannot improve the order of accuracy and is omitted.

[31] These algorithms cover several classes of numerical methods for Richards equation, including iterative and noniterative solvers, conservative and nonconservative formulations, as well as first- and second-order accurate approximations.

[32] The primary objectives of the noniterative scheme are (1) to maintain the proportionality between the user-prescribed tolerance and the actual temporal errors, and (2) to constrain the temporal errors in a reasonably uniform profile. The analysis focuses on temporal errors, since spatial errors arise due to the finite element approximation and are unrelated to the time stepping scheme. Therefore, unless specified otherwise, all solutions correspond to a uniform spatial mesh with 320 elements, \( \Delta z = 0.625 \text{ cm} \). This procedure isolates temporal errors and facilitates the analysis of time accuracy. Spatial errors are also briefly discussed to provide a more complete numerical assessment.

[33] The error measure used in this study is obtained by comparing the approximate and exact solutions at a series of a priori specified output times. In order to simplify the error analysis, absolute error requirements are enforced by setting \( \tau_R = 0 \) in the adaptive scheme. The error norm is defined as:

\[
\varepsilon_{EC}(\mathbf{t}^n, \tau_A) = \max_i |\psi_i^n - \overline{\psi}_i^n|,
\]

where \( \overline{\psi}_i^n \) is the “exact-in-time” solution for the 320 finite element grid, and \( i \) indexes the spatial nodes. We also consider \( \varepsilon_D(\mathbf{t}^n, \tau_A) = \max_i |\psi_i^n - \overline{\psi}_i^n| \), where \( \overline{\psi}_i^n \) is the solution obtained on a dense spatial mesh of 3200 elements. Since no closed-form analytic solution to this flow problem is available, the “exact” solutions were evaluated numerically by the noniterative adaptive scheme with a very tight error tolerance, \( \tau_A = 10^{-6} \text{ cm} \), and then verified using a range of error tolerances and the alternative numerical formulations. For the iterative schemes, we designate \( \tau_{ABS} \) as the absolute error tolerance of the Picard solver.

[34] In addition, the following measure of the mass balance error \( E_{MB} \) is used:

\[
E_{MB} = \max_n \left| 1 - \varepsilon_{MB}(\mathbf{t}^{n+1}) \right| \times 100\%,
\]

where \( \varepsilon_{MB}(\mathbf{t}^{n+1}) \) is the cumulative mass balance ratio at \( \mathbf{t}^{n+1} \), given by

\[
\varepsilon_{MB}(\mathbf{t}^{n+1}) = \frac{\sum_{i=1}^{N_e} (\theta_i^{n+1} - \theta_i^n) \Delta z}{\sum_{k=1}^{N_e} (q_{in,k}^{n} - q_{out,k}^{n}) \Delta t^k},
\]

where \( N_e \) is the number of spatial elements. The fluxes \( q_{in,k}^{n} \) are estimated using the backward rule \( q_{in,k}^{n} = q_{k+1}^{n} \) for the first-order schemes [Celio et al., 1990] and the trapezoidal rule \( q_{in,k}^{n} = \frac{1}{2}(q_1^{n} + q_{k+1}^{n}) \) for the second-order schemes. The fluxes \( q_1^{n} \) and \( q_{k+1}^{n} \) are obtained from the finite element equations.

[35] The measures of computational effort used in this paper are the number of time steps and the number of iterations per simulation. The number of time steps is important, since, as vividly illustrated below, the time step size may be strongly influenced by the convergence of the nonlinear solver. However, it is the total number of matrix formations and inversions that dominates the computational cost of implicit approximations, with negligible additional cost of coefficient linearization (12) and error control.

[36] The temporal error profiles of the time stepping algorithms with a range of a priori error tolerances are shown in Figure 2. Tables 1 and 2 contain the respective mass balance errors and run time statistics. Several conclusions regarding adaptive error control, noniterative linearizations, and order of accuracy issues for Richards equation can be drawn from these results.

[37] The adaptive error control allows all numerical algorithms to complete the simulation. The temporal error profiles are relatively uniform and proportional to the prescribed tolerance. As expected, the pressure-based algorithms have only minor mass balance errors (below 1%). Miller et al. [1998] reported that for a similar flow case, the fixed step modified Picard scheme did not converge for the range of time steps tested. Although this work shows that Picard iterations will converge provided the step size is reduced sufficiently, a fixed step algorithm would be extremely uneconomical for this problem, since the time step size necessary to maintain a consistent temporal accuracy in the entire solution varies by several orders of magnitude throughout the simulation.

[38] Tables 1 and 2 show that adaptive time stepping alone does not guarantee cost-effective solutions to this problem. Only the noniterative schemes produced economical solutions, while iterative methods suffered a severe loss of efficiency. For example, an accuracy of \( \pm 0.05 \text{ cm} \) requires 13,706 steps (16,060 iterations) with the second-order noniterative scheme, while 368,396 steps (almost 10,000,000 iterations) of the second-order iterative scheme lead to errors of \( \pm 0.1 \text{ cm} \). The CPU run times on a Pentium II 350 MHz are 1 min versus 7.4 hours!

[39] It can be seen from Table 1 that the second boundary condition period, where a portion of the domain becomes saturated, is by far the most difficult part of the simulation, accounting for \( \sim 75\% \) of the computational cost for the iterative schemes and the modified implicit Euler scheme and 30–60% for the new noniterative scheme. The third boundary condition period is also difficult, due to the transition of a section of the domain from saturated to unsaturated conditions, and accounts for 20–25% of total costs. The first time period, where the entire domain is unsaturated, is easy for all numerical schemes, requiring 10% of the cost for the noniterative schemes and \(< 0.05\% \) for the iterative schemes. The error versus cost analysis indicates that the noniterative schemes are slightly more efficient for the easy portions of the simulation (due to the lack of iterations) and much more efficient in the difficult time periods (due to better solver stability).

[40] The difficulties encountered by the iterative schemes arise due to ill-conditioned Jacobians of the discrete equations at \( \psi \approx 0 \). In this case, nonlinear iterations are only conditionally stable, stagnating or diverging unless stringent time step size constraints, much stricter than required for
temporal accuracy, are satisfied. A noniterative solver avoids this problem, as the iteration error growth is bounded by the single iteration and the solver linearization maintains the order of temporal accuracy. In addition, truncation error control is a useful safeguard against solver problems observed when the time step size is chosen without regard for the accuracy of the integration.

[41] The noniterative linearization (12) maintains the $O(\Delta t^2)$ accuracy of the time stepping scheme, as expected from the truncation error analysis (15). The noniterative adaptive scheme also retains the ability to constrain the temporal discretization errors near the user-prescribed tolerance. Since the temporal error profile is relatively uniform, it follows that the linearization (12) and (13) is compatible with the adaptive time approximation and error control.

[42] The modified implicit Euler scheme also benefits from truncation error control and is more efficient than both iterative schemes, but its first-order convergence makes the algorithm less efficient than the second-order noniterative scheme. This finding is in agreement with previous studies [Paniconi et al., 1991] and highlights the importance of adequate linearization in noniterative schemes: A linearization that reduces the order of accuracy of the time approximation also reduces the competitiveness of the algorithm. The investment in the $O(\Delta t^2)$ linearization is therefore justified.

**Figure 2.** Temporal error profiles for four time stepping schemes: (1) the new noniterative adaptive time stepping scheme (open, solid, and shot circles designate $\tau_A = 10^{-1}, 10^{-2},$ and $10^{-3}$ cm, respectively); (2) the modified implicit Euler scheme (open, solid, and shot squares designate $\tau_A = 10^{-2}, 10^{-4},$ and $10^{-6}$ cm, respectively); (3) the modified Picard scheme (open and solid triangles designate $\tau_A = 10^{-1}, \tau_{A(PIC)} = 10^{-4}$ cm and $\tau_A = 10^{-2}, \tau_{A(PIC)} = 10^{-4}$ cm, respectively); and (4) the standard Picard scheme (open and solid diamonds designate $\tau_A = 10^{-1}, \tau_{A(PIC)} = 10^{-4}$ cm and $\tau_A = 10^{-2}, \tau_{A(PIC)} = 10^{-4}$ cm, respectively).

**Table 1.** Performance Statistics for the Noniterative Time Stepping Schemes

<table>
<thead>
<tr>
<th></th>
<th>New $O(\Delta t^2)$ Noniterative Scheme</th>
<th>Modified Implicit Euler Scheme, $O(\Delta t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal tolerance $\tau_A$, cm</td>
<td>$10^{-1}$</td>
<td>$10^{-2}$</td>
</tr>
<tr>
<td>Total successful time steps</td>
<td>5,099</td>
<td>13,706</td>
</tr>
<tr>
<td>Time steps in T1</td>
<td>561</td>
<td>1,740</td>
</tr>
<tr>
<td>Time steps in T2</td>
<td>1,644</td>
<td>7,697</td>
</tr>
<tr>
<td>Time steps in T3</td>
<td>2,894</td>
<td>4,269</td>
</tr>
<tr>
<td>Failed steps</td>
<td>495</td>
<td>2,354</td>
</tr>
<tr>
<td>Total linear solutions</td>
<td>5,594</td>
<td>16,060</td>
</tr>
<tr>
<td>CPU run time, min</td>
<td>0.30</td>
<td>0.9</td>
</tr>
<tr>
<td>Time error, max$(\varepsilon(t))$, cm</td>
<td>2.5</td>
<td>$3.7 \times 10^{-2}$</td>
</tr>
<tr>
<td>Mass balance error $E_{MB}$, %</td>
<td>0.15</td>
<td>0.0047</td>
</tr>
</tbody>
</table>

* T1, T2, and T3 refer to the three time periods within the simulation, equation (22). CPU times are listed for a Pentium II 350 MHz processor. Each time step requires a single linear system solution using tridiagonal LU decomposition.
In contrast to the noniterative schemes, none of the iterative Picard algorithms is competitive for simulating saturated conditions in this soil, since failures of the nonlinear solver to converge (solver failures in Table 2) dramatically reduce the time step size irrespective of temporal truncation error requirements. Ironically, the accuracy benefits of smaller step sizes are lost, since whenever acceptable solver performance is possible, the time step is increased. This behavior explains why the error of iterative schemes is relatively large despite an enormous number of time steps; global errors grow immediately after a few time steps that are chosen according to the prescribed user tolerance. It is also noted that the mass conservative modified Picard scheme has no accuracy or efficiency advantage over the formally nonconservative algorithms; all iterative schemes are severely affected by the lack of smoothness near \( \psi \approx 0 \).

Although the nonsmooth nonlinearities at \( \psi \approx 0 \) lead to a relatively frequent time step rejection in noniterative schemes (~12% for the \( O(\Delta t^2) \) scheme and ~20% for the \( O(\Delta t) \) scheme), the cheap cost per step of these methods more than compensates for these failures. Furthermore, the apparently small number of failed steps in the iterative schemes is deceptive, since it is masked by the massive failure rate of the nonlinear solver, which failed to converge in ~30% of the time step attempts. These solver failures also reduce the benefits of second-order accuracy of the adaptive scheme with the standard Picard solver.

It can be seen that the proposed adaptive \( O(\Delta t^2) \) noniterative scheme is the most efficient temporal integration method for this test problem, followed by the modified implicit Euler scheme with formal truncation error control. Adaptive implementations of the standard Picard scheme and the mass conservative modified Picard scheme are able to complete the simulation, but suffer a severe loss of computational efficiency when a part of the soil domain becomes saturated.

Although adequate temporal integration is essential for a successful solution of Richards equation, it is insufficient, since the overall accuracy also depends on the spatial approximation. In general, it is futile to refine the time discretization unless the spatial mesh is also improved. Figure 3 shows that when a 320-element spatial mesh is used, spatial errors limit the global accuracy of simulations when temporal error tolerances below ±0.1 cm are imposed.

### Table 2. Performance Statistics for the Iterative Time Stepping Schemes

<table>
<thead>
<tr>
<th></th>
<th>Modified Picard Scheme, ( O(\Delta t) )</th>
<th>Standard Picard Scheme, ( O(\Delta t^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal tolerance ( \tau_t ), cm</td>
<td>( 10^{-1} )</td>
<td>( 10^{-2} )</td>
</tr>
<tr>
<td>Iteration tolerance ( \tau_n ), cm</td>
<td>( 10^{-1} )</td>
<td>( 10^{-1} )</td>
</tr>
<tr>
<td>Total successful time steps</td>
<td>354,577</td>
<td>435,135</td>
</tr>
<tr>
<td>Time steps in T1</td>
<td>547</td>
<td>1,726</td>
</tr>
<tr>
<td>Time steps in T2</td>
<td>262,168</td>
<td>311,786</td>
</tr>
<tr>
<td>Time steps in T3</td>
<td>91,862</td>
<td>121,623</td>
</tr>
<tr>
<td>Failed steps</td>
<td>33</td>
<td>62</td>
</tr>
<tr>
<td>Total linear solutions</td>
<td>9,623,026</td>
<td>11,693,094</td>
</tr>
<tr>
<td>Linear solutions in T1</td>
<td>2,072</td>
<td>4,862</td>
</tr>
<tr>
<td>Linear solutions in T2</td>
<td>7,107,106</td>
<td>8,420,624</td>
</tr>
<tr>
<td>Linear solutions in T3</td>
<td>2,513,848</td>
<td>3,267,608</td>
</tr>
<tr>
<td>Solver failures</td>
<td>182,882</td>
<td>222,112</td>
</tr>
<tr>
<td>CPU run time, min</td>
<td>425</td>
<td>595</td>
</tr>
<tr>
<td>Time error, max(( e_{t} )), cm</td>
<td>1.2</td>
<td>0.38</td>
</tr>
<tr>
<td>Mass balance error ( E_{\text{mb}} ), %</td>
<td>(&lt;10^{-6})</td>
<td>(&lt;10^{-6})</td>
</tr>
</tbody>
</table>

![Figure 3](image-url). Total error profiles for the noniterative adaptive time stepping schemes (open and solid circles represent \( \tau_t = 10^{-1} \) cm and \( \tau_t = 10^{-2} \) cm, respectively) compared with an exact-in-time solution (open squares) and a solution with the noniterative scheme and fixed time steps of size \( \Delta t = 30 \) sec (diamonds).
since refinement of the temporal error tolerance $\tau_d$ does not lead to a reduction in the global solution error. As the solution approaches steady state, the spatial errors decrease dramatically. The apparently inconsistent result at $t = 200,000$ s, where a coarser temporal tolerance leads to higher overall error, is due to a cancellation of a portion of the spatial error by temporal error.

The error profiles of the adaptive noniterative scheme in Figure 3 are almost indistinguishable from the exact-in-time solution, implying that no further improvement in accuracy is possible unless the spatial approximation is refined. Conversely, the error of the fixed step noniterative scheme is considerable larger and would remain large even if a very dense spatial grid were used. It is also noted that fixed step standard and modified Picard schemes did not converge unless extremely small time steps $\Delta t \ll 1$ s were used.

Moreover, an examination of the spatial pressure profiles in this flow case shows that when uniform spatial grids are used, only about 10% of the nodes are located in the highly nonlinear infiltration front region. The remaining spatial nodes contribute very little to accuracy, yet increase the computational load by a factor of 10 or more. Therefore we expect that discretization algorithms that are locally adaptive in both time and space would be a fundamental advance in numerical methods for Richards equation and similar nonlinear transport PDEs.

### 3.2. Test Problem 2

A second test study compares the $O(\Delta t^2)$ noniterative algorithm to the published DASPK-KAM solution of Miller et al. [1998]. The $K$ arithmetic mean (KAM) algorithm was chosen since the arithmetic mean of elemental conductivity is used in the time stepping algorithms presented here. The material properties used in the first test problem are prescribed in a 200-cm vertical soil column discretized into 320 elements. Boundary conditions are $\psi(z = 200, t) = 0$ and $\psi(z = 0, t) = 10$ cm. A hydrostatic initial condition is employed with $\psi(z = 200, t = 0) = 0$. This is the most difficult problem considered by Miller et al. [1998].

Solutions are compared at $t = 1$ day with (1) a dense spatial-temporal grid solution $\bar{\psi}$ computed with $N_e = 10,240$ elements and $\tau_d = 10^{-5}$ cm; and (2) an exact-in-time solution $\psi$ obtained on the 320-element grid with $\tau_d = 10^{-3}$ cm. For the dense spatial-temporal solution an error measure is defined as

$$\|\varepsilon_D\|_2 = \sqrt{\frac{1}{N_e + 1} \sum_{j=1}^{N_e+1} (\bar{\psi}_j - \psi_j)^2}.$$  

An analogous root-mean-square error $\|\varepsilon_C\|_2$ is defined with respect to the exact-in-time solution $\psi$. The comparison with $\bar{\psi}$ yields the combined spatial-temporal errors, whereas the comparison with $\psi$ yields the temporal errors only.

Table 3. Comparison of the Adaptive Noniterative Scheme With the Published Performance of DASPK [Miller et al., 1998, Table 4]

<table>
<thead>
<tr>
<th>Error Tolerance, cm</th>
<th>Number of Linear Solutions</th>
<th>Error $|\varepsilon_D|_2$, cm</th>
<th>Error $|\varepsilon_C|_2$, cm</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_d = 1$</td>
<td>1,517</td>
<td>8.72</td>
<td>9.32</td>
</tr>
<tr>
<td>$\tau_d = 5 \times 10^{-1}$</td>
<td>2,322</td>
<td>3.40</td>
<td>4.14</td>
</tr>
<tr>
<td>$\tau_d = 1 \times 10^{-1}$</td>
<td>2,086</td>
<td>7.77 x 10^{-1}</td>
<td>3.52 x 10^{-2}</td>
</tr>
<tr>
<td>$\tau_d = 5 \times 10^{-2}$</td>
<td>2,916</td>
<td>8.10 x 10^{-1}</td>
<td>2.68 x 10^{-2}</td>
</tr>
<tr>
<td>$\tau_d = 1 \times 10^{-2}$</td>
<td>11,340</td>
<td>8.09 x 10^{-1}</td>
<td>2.38 x 10^{-3}</td>
</tr>
<tr>
<td>$\tau_d = 1 \times 10^{-3}$</td>
<td>52,971</td>
<td>8.10 x 10^{-1}</td>
<td>8.80 x 10^{-4}</td>
</tr>
</tbody>
</table>

DASPK-KAM [Miller et al., 1998]

| Variance = 1 | 1,506 | 1.48 x 10^{-2} | N/A |
| Variance = 5 x 10^{-1} | 578 | 5.93 x 10^{-2} | N/A |
| Variance = 1 x 10^{-1} | 307,120 | 2.11 x 10^{-2} | N/A |
| Variance = 5 x 10^{-2} | 151,157 | 3.27 x 10^{-2} | N/A |
| Variance = 1 x 10^{-2} | 85,623 | 2.39 x 10^{-2} | N/A |

In contrast, the published performance of DASPK-KAM does not display the expected convergence behavior as the time discretization is refined and computational effort increased: Even after more than 50,000 linear solutions the truncation errors do not consistently converge to spatial errors. Finally, Table 3 shows an apparent difference in the magnitude of the errors of the two schemes. This difference could be explained by the following: (1) temporal errors in the surrogate “exact” solution used by Miller et al., which was computed using a fixed step modified Picard scheme without explicit temporal error control, or (2) differences in the spatial approximations used.

The difference between the reported errors and convergence results for the noniterative and DASPK-KAM algorithms makes it difficult to conclusively determine the relative accuracy and efficiency of the two schemes. However, while the new adaptive algorithm converges quickly to the spatial errors, the same conclusion cannot be made for the DASPK-KAM solution.

While the results presented here do not cover all the many numerical methods available for Richards equation, additional comments can be made about some alternatives. Newton methods are typically more robust than Picard iterations; however, the nonsmoothness of the constitutive functions near $\psi = 0$ will also severely affect these methods [Miller et al., 1998]. In some Richards equation applications the Newton scheme requires such small time steps that it is even less practical than the Picard method [Tocci et al., 1997]. Using alternative elemental conductivity estimates and smoothing the constitutive functions with Hermite splines improve the performance of DASPK [Miller et al., 1998] and would also benefit the noniterative scheme. This work, however, demonstrates that accurate and cost-effective solutions can also be obtained with the much more widely used arithmetic mean estimates of elemental conductivities.
4. Conclusions

[55] The study explores the performance of noniterative adaptive algorithms for the time integration of Richards equation and introduces a new second-order accurate noniterative adaptive formulation. Accurate noniterative linearizations avoid numerical divergence problems that hinder common iterative solvers when modeling saturated soils with nonsmooth constitutive functions, while the lack of iterations makes noniterative formulations more efficient than analogous time stepping schemes with iterative solvers. The reliability and efficiency of the noniterative schemes is further improved by adaptive truncation error control.

[56] Numerical assessment shows that (1) noniterative schemes for Richards equation are an attractive alternative to conventional iterative schemes, allowing accurate and cost-effective solutions of problems that could not be satisfactorily handled using standard algorithms, (2) the new second-order noniterative linearization is more efficient than first-order approximations and is competitive with the variable-order variable-step DASPK-KAM algorithm, and (3) adaptive time step variation is particularly important for the solution of Richards equation in variably saturated soils.

[57] The noniterative formulation is conceptually simple and can be incorporated into existing backward Euler codes. It could also be employed for nonlinear differential equations in other areas of applied science and engineering. Since adaptive time approximations are formulated for arbitrary spatial discretizations, while noniterative linearizations could also be implemented in alternative time stepping schemes, the development of iterative and noniterative algorithms that are locally adaptive in both space and time is a promising direction for future research.

References


