WASY Software

FEFLOW®

Finite Element Subsurface Flow
& Transport Simulation System

White Papers
Vol. I

WASY GmbH
Institute for
Water Resources Planning
and Systems Research

Contents

1. On the primary variable switching technique for simulating unsaturated-saturated flows 9

2. Treatment of free surfaces in 2D and 3D groundwater modeling 67

Contents

8. An efficient method for computing groundwater residence times. 139

9. Discrete feature modeling of flow, mass and heat transport processes by using FEFLOW 149

Contents

14. Extended formulations of constraints for Cauchy-type (3rd kind) boundary conditions in FEFLOW. 269

15. Nonlinear dispersion in density-dependent mass transport 275

16. Consistent velocity approximation in the finite-element simulation of density-dependent mass and heat transport processes 281

Contents

17. Coupled groundwater flow and transport: Thermohaline and 3D convection systems. 313

Subject Index 357

Author Index 365

On the primary variable switching technique for simulating unsaturated-saturated flows

1

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

[a]*WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany*
[b]*Centre d'Hydrogéologie, Université de 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.

Key words: unsaturated-saturated flow, primary variable switching, Newton technique, finite elements, time stepping control, benchmarking, capillary barrier

1.1 Introduction

In the modeling of unsaturated-saturated flow processes several alternatives exist for numerically solving the governing balance equations with their nonlinear 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 ψ 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 modeler 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 (ψ-) form, where the primary variable is the pressure head

1. On the primary variable switching technique for simulating unsaturated-saturated flows

(or the hydraulic head), (2) the saturation-based (s-) form, where the saturation (or the moisture content θ) 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 ψ-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[15,18,22,23,30,32,33,36,37,39,43,44]. But, it has been shown[1,4,31,45] that the ψ-based form can produce significant global mass balance errors unless very small time steps are used. The ψ-based approach can be 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[38]. 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, chord 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[37].

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.[42], Vogel et al.[47] and Ju and Kung[25] for different situations. Fuhrmann[16] and Lehmann and Ackerer[28] enhanced the mixed form by using a Newton iterative scheme instead of the Picard iteration. Lehmann and Ackerer[28] 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. Additionally, the common transformation into the s-form (and the equivalent water-content-based form) of the Richards equation is restricted to homogeneous soils as thoroughly discussed by LaBolle and Clausnitzer[27]. Note further 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 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.

Notice, the common saturation-based and the equivalent water-content-based forms of the Richards equation are restricted to unsaturated flow conditions (s < 1) and homogeneous soils (parameters of unsaturated soil properties must be spatially invariant).

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[35]. 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 for heterogeneous porous media. 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[44], Kirkland et al.[26], Forsyth et al.[13], Webb[48], and Forsyth and Kropinski[14]. Further comparative studies for finding the 'best' solution strategy in practical modeling of unsaturated-saturated flows are required.

1.2 Basic Equations

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

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

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

$$\boldsymbol{q} = -K_r(s)\boldsymbol{K}(\nabla h + \chi\boldsymbol{e}) = -K_r(s)\boldsymbol{K}[\nabla\psi + (1+\chi)\boldsymbol{e}] \quad \text{(1-2)}$$

In eqns (1-1) and (1-2),

$h = \psi + z$, hydraulic (piezometric) head;

1. On the primary variable switching technique for simulating unsaturated-saturated flows

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);

$\boldsymbol{q}$ 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;

ε porosity;

γ fluid compressibility;

Υ coefficient of skeleton compressibility;

$K_r(s)$ relative hydraulic conductivity ($0 < K_r \leq 1$, $K_r = 1$ if saturated at $s = 1$);

$\boldsymbol{K}$ tensor of hydraulic conductivity for the saturated medium (anisotropy);

χ 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 ψ, as well as its inverse, the pressure head ψ as a function of the saturation s, and (2) for the relative hydraulic conductivity K_r as a function of either the pressure head ψ or the saturation s. The following empirical relationships are used for the present study[2,47]:

van Genuchten-Mualem parametric model:

$$s_e = \begin{cases} \dfrac{1}{\left[1 + |\alpha\psi|^n\right]^m} & \text{for} \quad \psi < \psi_a \\ 1 & \text{for} \quad \psi \geq \psi_a \end{cases} \tag{1-3}$$

$$K_r = s_e^{\frac{1}{2}}\left\{1 - \left[1 - s_e^{\frac{1}{m}}\right]^m\right\}^2 \tag{1-4}$$

Brooks-Corey parametric model:

$$s_e = \begin{cases} \dfrac{1}{|\alpha\psi|^n} & \text{for} \quad \psi < -1/\alpha \\ 1 & \text{for} \quad \psi \geq -1/\alpha \end{cases} \tag{1-5}$$

$$K_r = s_e^{\kappa} \tag{1-6}$$

with the effective saturation

$$s_e = \frac{s - s_r}{s_s - s_r} \tag{1-7}$$

in which

s_e effective saturation;

s_r residual saturation;

s_s maximum saturation;

ψ_a air-entry pressure head, $\psi_a \leq 0$;

α curve-fitting parameter;

n pore size distribution index, $n \geq 1$;

$m = 1 - 1/n$, curve fitting parameter (Mualem assumption);

$\kappa = 2/n + l + 2$, curve-fitting parameter;

l pore-connectivity parameter;

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

$$R(s, h) = S_o \cdot s(\psi)\frac{\partial\psi}{\partial t} + \varepsilon\frac{\partial s(\psi)}{\partial t} - $$
$$-\nabla \cdot \{K_r(s)\boldsymbol{K}[\nabla\psi + (1 + \chi)\boldsymbol{e}]\} - Q = 0 \tag{1-8}$$

which has to be solved either for ψ (and h) or s. The retention curves (1-3) or (1-5) can be used to convert one variable to the other (and *vice versa*), *viz.*,

$$s = f(\psi)$$
$$\psi = f^{-1}(s) \tag{1-9}$$

1.3 Finite Element Formulation

Let $\Omega \subset \mathfrak{R}^D$ 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 Γ denote the boundary of Ω, the weak form of the mass balance equation (1-1) can be written as

$$\int_\Omega w S_o s(\psi) \frac{\partial \psi}{\partial t} + \int_\Omega w \varepsilon \frac{\partial s}{\partial t} - \int_\Omega q \cdot \nabla w = \int_\Omega w Q - \int_\Gamma w q_n \tag{1-10}$$

and with eqn (1-2) as

$$\int_\Omega w S_o s(\psi) \frac{\partial \psi}{\partial t} + \int_\Omega w \varepsilon \frac{\partial s}{\partial t} + \int_\Omega \nabla w \cdot [K_r(s) K \cdot \nabla \psi] =$$
$$= \int_\Omega w Q - \int_\Gamma w q_n - \int_\Omega \nabla w \cdot [K_r(s) K \cdot (1 + \chi) e] \tag{1-11}$$

where w is a test function and q_n corresponds to the normal fluid flux directed positive outward on Γ.

In the finite element context a spatial semi-discretization Ω^h of the continuum domain Ω is achieved by the union of a set of nonoverlapping subdomains Ω_e,

the finite elements, as

$$\Omega \approx \Omega^h \equiv \bigcup_e \Omega_e \tag{1-12}$$

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

$$\psi(x_i, t) \approx \psi^h(x_i, t) = N_I(x_i)\psi_I(t)$$
$$s(x_i, t) \approx s^h(x_i, t) = N_I(x_i)s_I(t) \tag{1-13}$$

and, respectively,

$$K_r(x_i, t) \approx K_r^h(x_i, t) = N_I(x_i)K_{rI}(t) \tag{1-14}$$

where $i = 1, ..., D$ represents coordinate indices, $I = 1, ..., M$ 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 coordinates. 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, eqn (1-11) leads to the following global matrix system of M equations

$$O(s) \cdot \dot{\Psi} + B \cdot \dot{s} + K(s) \cdot \Psi - F(s) = 0 \tag{1-15}$$

with its components written in indicial notation

$$O_{IJ}(s) = \sum_e \int_{\Omega_e} N_I S_o s(\psi) \delta_{IJ} \tag{1-16a}$$

$$B_{IJ} = \sum_e \int_{\Omega_e} N_I \varepsilon \delta_{IJ} \tag{1-16b}$$

$$K_{IJ}(s) = \sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i} K_r(s) K_{ij} \frac{\partial N_J}{\partial x_j} \tag{1-16c}$$

$$F_I(s) = \sum_e \int_{\Omega_e} N_I Q - \sum_e \int_{\Gamma_e} N_I q_n$$
$$- \sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i} K_r(s) K_{ij} (1 + \chi) e_j \tag{1-16d}$$

where the subscripts $I, J = 1, \dots, M$ denote nodal indices, $i, j = 1, \dots, D$ are spatial indices of the Cartesian coordinates, and δ_{IJ} is the Kronecker operator. The superposed dot means differentiation with respect to time t. Nonlinearities 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 (1-15) is highly nonlinear due to the functional dependence of the constitutive relationships (1-3)-(1-6) for the saturation and the relative conductivity.

The discretized form (1-15) of the Richards equation is based on the mixed formulation (1-8), where the fluid and medium compressibility S_o relates to the pressure head ψ. 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 (1-15) is unconditionally mass-conservative for a vanishing S_o-term only.

1.4 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 nondissipative trapezoid rule (TR) with a second-order accuracy are enforced.

Denoting the time plane by the superscript n, the implicit form of eqn (1-15) reads

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

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

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

and for the TR scheme, by

$$\dot{\Psi}^{n+1} = \frac{2}{\Delta t_n}(\Psi^{n+1} - \Psi^n) - \dot{\Psi}^n$$

$$\dot{s}^{n+1} = \frac{2}{\Delta t_n}(s^{n+1} - s^n) - \dot{s}^n$$

(1-19)

Inserting eqns (1-18) and (1-19) into eqn (1-17) results in

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

(1-20)

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 eqn (1-20) 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 $\dot{\Psi}^n$ and $\dot{s}^n$ at the previous time plane, and by the factor $2/\Delta t_n$ instead of $1/\Delta t_n$.

1.5 Primary Variable Switching Methodology

To solve the basic matrix system (1-20) 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' ψ-, 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 ψ_I or s_I. Accordingly, we can consider X as a vector containing the different primary variables in the solution space Ω^h as

$$X \in (\Psi, s)$$

(1-21)

Hence, the matrix system (1-20) can be written in the form

$$R^{n+1}(X) = 0$$

(1-22)

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

The solution of the nonlinear equations (1-22), i.e., the vector of primary variables X, is performed by the Newton method, *viz.*,

$$J^X(\Psi_\tau^{n+1}, s_\tau^{n+1})\Delta X_\tau^{n+1} = -R_\tau^{n+1}(\Psi, s)$$

(1-23a)

with the increment

$$\Delta X_\tau^{n+1} = X_{\tau+1}^{n+1} - X_\tau^{n+1}$$

(1-23b)

and the Jacobian J^X expressed in indicial notation as

1. On the primary variable switching technique for simulating unsaturated-saturated flows

Notice, *the differentiation in form of (1-23c) does not restrict the approach to invariant soil properties, because there is never a derivation with respect to a spatial coordinate. This is a distinct advantage of the variable switching technique, which can solve the problem even in the saturation variable (if s < 1) for heterogeneous soil parameters. In contrast, common formulations in the saturation (or water-content) form are restricted to homogeneous soil characteristics[27].*

$$J_{IJ}^{X}(\Psi_\tau^{n+1}, s_\tau^{n+1}) = \frac{\partial R_I^{n+1}(\Psi_\tau^{n+1}, s_\tau^{n+1})}{\partial X_{\tau J}^{n+1}} \tag{1-23c}$$

where τ denotes the iteration number.

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

IF ($s_{\tau I}^{n+1} \geq tol_f$) THEN

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

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

ELSE IF ($s_{\tau I}^{n+1} < tol_b$) THEN

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

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

ELSE

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

ENDIF

The Newton approach requires continuous derivatives of the Jacobians J^{ψ} and J^{s} with respect to the pressure head ψ and the saturation s, respectively. In the present finite element method the variables ψ and s are approximated in a continuous manner according to (1-13) 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 (1-3) or (1-5) 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 Diersch and Kolditz[10].

The switching tolerances tol_f and tol_b have to be appropriately chosen. The following requirements are necessary

$$tol_f < 1 \qquad tol_f \neq tol_b \tag{1-26}$$

The Jacobians $\boldsymbol{J}^{X}$ can be computed either numerically or analytically. The analytical method is more efficient[28] 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 $\boldsymbol{J}^{\psi}(\Psi_\tau^{n+1}, s_\tau^{n+1})$ of eqn (1-24) and $\boldsymbol{J}^{s}(s_\tau^{n+1}, \Psi_\tau^{n+1})$ of eqn (1-25) are summarized in the Appendices A and B, respectively. Otherwise, the residual $R_{\tau I}^{n+1}(\Psi, s)$ at the iterate τ and node I is independent of the actually used primary variables X_I and is computed according to eqn (1-20) in the following way

$$-R_{\tau I}^{n+1}(\Psi, s) = -\left(\frac{\sigma O_{IJ}(s_\tau^{n+1})}{\Delta t_n} + K_{IJ}(s_\tau^{n+1})\right) \cdot \psi_{\tau J}^{n+1}$$

$$- \frac{\sigma B_{IJ}}{\Delta t_n} \cdot s_{\tau J}^{n+1}$$

$$+ O_{IJ}(s_\tau^{n+1}) \cdot \left(\frac{\sigma}{\Delta t_n}\psi_J^n + (\sigma-1)\dot\psi_J^n\right)$$

$$+ B_{IJ} \cdot \left(\frac{\sigma}{\Delta t_n}s_J^n + (\sigma-1)\dot s_J^n\right) + F_I(s_\tau^{n+1})$$

$$(1\text{-}27)$$

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 J^X. Traditionally, the assembling process is performed by

$$J_{IJ}^X = \sum_e \int_{\Omega_e} \{\dots\}_{\forall I, \forall J} \qquad (1\text{-}28)$$

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

$$J_{IJ}^X = \sum_I \sum_{e \in \eta_I} \int_{\Omega_e} \{\dots\}_{I, \forall J} \qquad (1\text{-}29)$$

where the contributions from an adjacent element patch η_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 pre-

vious solution strategies are encompassed as special cases. Taking the pressure head ψ as primary variable, omitting for simplicity the compressibility term $O(S_o)$ and considering only the fully implicit BE scheme, we obtain from eqns (1-24) and (A1)

$$\left(K + \Psi_\tau^{n+1}\frac{\partial K(s_\tau^{n+1})}{\partial \Psi_\tau^{n+1}} + \frac{B}{\Delta t_n}\frac{\partial s_\tau^{n+1}}{\partial \Psi_\tau^{n+1}} - \frac{\partial F(s_\tau^{n+1})}{\partial \Psi_\tau^{n+1}}\right)(\Psi_{\tau+1}^{n+1} - \Psi_\tau^{n+1}) =$$

$$= -K\Psi_\tau^{n+1} - \frac{B}{\Delta t_n}(s_\tau^{n+1} - s^n) + F(s_\tau^{n+1})$$

$$(1\text{-}30)$$

which is the Newton scheme of the mixed $\psi - s$-form of the Richards equation[16,28]. Furthermore, the modified Picard scheme for the mixed $\psi - s$-form of the Richards equation[4] can be deduced from eqn (1-30) by dropping the partial Jacobians of the 2nd and 4th term of the left-hand side of eqn (1-30), yielding

$$\left(K + \frac{B}{\Delta t_n}C_\tau^{n+1}\right)\Psi_{\tau+1}^{n+1} = \frac{B}{\Delta t_n}C_\tau^{n+1}\Psi_\tau^{n+1} - \frac{B}{\Delta t_n}(s_\tau^{n+1} - s^n) + F(s_\tau^{n+1})$$

$$(1\text{-}31)$$

with the moisture capacity (A7) $C_\tau^{n+1} = \partial s_\tau^{n+1}/\partial \Psi_\tau^{n+1}$. Finally, the common ψ-based form is easily obtained from eqn (1-31) if the saturation terms of the right-hand side are expressed by their derivatives with respect to the pressure head: $s_\tau^{n+1} - s^n = C_\tau^{n+1}(\Psi_\tau^{n+1} - \Psi^n)$

$$\left(K + \frac{B}{\Delta t_n}C_\tau^{n+1}\right)\Psi_{\tau+1}^{n+1} = \frac{B}{\Delta t_n}C_\tau^{n+1}\Psi^n + F(s_\tau^{n+1}) \quad (1\text{-}32)$$

While the Newton scheme applied to the primary variable switching technique in eqns (1-24) and (1-25) and to the mixed form (1-30) is quadratically convergent, the Picard-type solutions (1-31) and (1-32) provide only a linearly convergent accuracy. One notes here that the matrix systems of the Newton method (1-24), (1-25) and (1-30) are always unsymmetric, while the Picard schemes in eqns (1-31) and (1-32) 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.*[36], Paniconi and Putti[37] and Miller *et al.*[30] who started from a ψ -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.

1.6   Solution Control

1.6.1   Adaptive predictor-corrector one-step Newton (PCOSN) time marching scheme

Generally, the control of the solution of the resulting highly nonlinear matrix systems (1-24) and (1-25) 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. 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 prob-

lems. 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 δ can also be used as the Newton convergence criterion for the iterate τ. This is called the one-step Newton method where δ 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 $\dot{s}^0$ from eqn (1-17) as

$$[O(s^0) + BC^0] \cdot \dot{\Psi}^0 = -K(s^0) \cdot \Psi^0 + F(s^0) \qquad (1\text{-}33)$$

and with

$$\dot{s}^0 = C^0 \cdot \dot{\Psi}^0 \qquad (1\text{-}34)$$

where C^0 is the initial moisture capacity vector according to eqn (A7), Ψ^0 and s^0 are the initial distributions of the pressure head ψ and the saturation s, respectively. Furthermore, we choose an initial time step size Δt_0.

STEP 1: Predictor solutions

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

$$X_p^{n+1} = X^n + \Delta t_n \dot{X}^n \qquad (1\text{-}35)$$

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

$$X_p^{n+1} = X^n + \frac{\Delta t_n}{2}\left[\left(2 + \frac{\Delta t_n}{\Delta t_{n-1}}\right)\dot{X}^n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{X}^{n-1}\right] \qquad (1\text{-}36)$$

Note here that, since $\dot{X}^{n-1}$ 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 $\dot{X}^0$ is available from eqns (1-33) and (1-34). 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 nonlinear matrix equations (1-24) and (1-25) are linearized by using the corresponding predictors. Accordingly, the Newton iterates are taken as

$$\Psi_\tau^{n+1} = \Psi_p^{n+1} \qquad s_\tau^{n+1} = s_p^{n+1} \qquad (1\text{-}37)$$

STEP 2: Corrector solutions

Depending on the primary variable switching criteria stated above the following matrix systems (1-24), (1-25) arise

$$J_{IJ}^\psi(\Psi_p^{n+1}, s_p^{n+1})\Delta\psi_{pJ}^{n+1} = -R_{pI}^{n+1}(\Psi, s)$$
$$\Delta\psi_{pJ}^{n+1} = \psi_J^{n+1} - \psi_{pJ}^{n+1} \qquad (1\text{-}38)$$

to solve the pressure head Ψ^{n+1} or

$$J_{IJ}^s(\Psi_p^{n+1}, s_p^{n+1})\Delta s_{pJ}^{n+1} = -R_{pI}^{n+1}(\Psi, s)$$
$$\Delta s_{pJ}^{n+1} = s_J^{n+1} - s_{pJ}^{n+1} \qquad (1\text{-}39)$$

to solve the saturation s^{n+1}, where the (predicted) residual R_p^{n+1} in eqns (1-38) and (1-39) is also evaluated by using the predictor solutions Ψ_p^{n+1} and s_p^{n+1} applied to the τ-terms in eqn (1-27). Note that the predictor of the FE (1-35) is used for the BE ($\sigma = 1$) and that the predictor of the AB (1-36) is used for the TR ($\sigma = 2$) in eqns (1-38) and (1-39). 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 $\dot{X}^{n+1}$ are computed for the FE

$$\dot{X}^{n+1} = \frac{1}{\Delta t_n}(X^{n+1} - X^n) \qquad (1\text{-}40)$$

by using the BE (1-18) and for the AB

$$\dot{X}^{n+1} = \frac{2}{\Delta t_n}(X^{n+1} - X^n) - \dot{X}^n \qquad (1\text{-}41)$$

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

by modifying the TR (1-19) according to Bixler[3].

STEP 4: Error estimation

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

$$d^{n+1} = \phi(X^{n+1} - X_p^{n+1}) \qquad (1\text{-}42a)$$

with

$$\phi = \begin{cases} \dfrac{1}{2} & \text{for FE/BE} \\[2ex] \dfrac{1}{3\left(1 + \dfrac{\Delta t_{n-1}}{\Delta t_n}\right)} & \text{for AB/TR} \end{cases} \qquad (1\text{-}42b)$$

Appropriate error norms are applied for the vector

d^{n+1}. Commonly, the weighted RMS L_2 error norm

$$\|d^{n+1}\|_{L_2} = \left[\frac{1}{M}\left(\sum_I^M \left|\frac{d_I^{n+1}}{X_{max}^{n+1}}\right|^2\right)\right]^{1/2} \qquad (1\text{-}43)$$

and the maximum L_∞ error norm

$$\|d^{n+1}\|_{L_\infty} = \frac{1}{X_{max}^{n+1}}\max_I \left|d_I^{n+1}\right| \qquad (1\text{-}44)$$

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 (1-42a), (1-43), (1-44), the current time step size Δt_n, and a user-specified error tolerance δ as[17]

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

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

1. If

$$\Delta t_{n+1} \geq \Delta t_n \qquad (1\text{-}46a)$$

the current solution X^{n+1} is accurate within the error bound defined by δ and the increase of the time step is always accepted.

2. If

$$\zeta \Delta t_n \leq \Delta t_{n+1} < \Delta t_n \qquad (1\text{-}46b)$$

where ζ is typically 0.85, the solution X^{n+1} is accepted but the time step size is not changed, i.e., $\Delta t_{n+1} = \Delta t_n$.

3. If

$$\Delta t_{n+1} < \zeta \Delta t_n \qquad (1\text{-}46c)$$

the solution X^{n+1} cannot be accepted within the required error tolerance δ and has to be rejected. The proposed new time step size (1-45) is reduced according to[5]

$$\Delta t_n^{reduced} = \frac{\Delta t_n^2}{\Delta t_{n+1}} \left(\frac{\delta}{\left\| d^{n+1} \right\|_{L_p}} \right) \qquad (1\text{-}46d)$$

and the solution is repeated for the time plane $n+1$ with $\Delta t_n = \Delta t_n^{reduced}$.

It is important to note that the error tolerance δ is the only user-specified parameter to control the entire solution process. The starting-up phase is still influenced by the initial time step Δt_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., $\Delta t_n \leq \Delta t^{max}$. Sec-

ondly, the rate for changing the time step size $\Xi = \Delta t_{n+1} / \Delta t_n$ has also to be limited, i.e., $\Xi \leq \Xi_{max}$ (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 (1-38) or (1-39) are solved via a BiCGSTAB iterative solver[46] preconditioned by an incomplete Crout decomposition scheme. The preconditioning process automatically provides a suited scaling of the final matrix system. Otherwise, taking the predictor solutions (1-35) or (1-36) the derivative terms (A7) and (B7), namely the moisture capacity and inverse moisture capacity terms, respectively, are easily computed by chord 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 δ. At the same time, δ 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.

1.6.2 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 Δt_n at time plane n (at initial time we start with a sufficiently small Δt_0) we solve for the new Newton iteration $\tau + 1$ either

$$J^\psi(\Psi_\tau^{n+1}, s_\tau^{n+1})\Delta\Psi_\tau^{n+1} = -R_\tau^{n+1}(\Psi, s)$$
$$\Delta\Psi_\tau^{n+1} = \Psi_{\tau+1}^{n+1} - \Psi_\tau^{n+1} \tag{1-47}$$

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

$$J^s(\Psi_\tau^{n+1}, s_\tau^{n+1})\Delta s_\tau^{n+1} = -R_\tau^{n+1}(\Psi, s)$$
$$\Delta s_\tau^{n+1} = s_{\tau+1}^{n+1} - s_\tau^{n+1} \tag{1-48}$$

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

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

with

$$d_\tau^{n+1} = X_{\tau+1}^{n+1} - X_\tau^{n+1} \tag{1-49b}$$

and where $\left\|d_\tau^{n+1}\right\|_{L_p}$ can be used as a RMS ($p = 2$, eqn

(1-43)) or maximum ($p = \infty$, eqn (1-44)) error norm.

STEP 2: Tactic of time stepping at successful Newton convergence
If Newton iterations have converged a new provisional step size Δt_{n+1} can be computed in the following way:

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

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

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

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

STEP 3: Tactic of time stepping if Newton iteration fails
The convergence criterion for the Newton method is given by eqn (1-49a). If the Newton scheme does not converge within a maximum number of nonlinear iterations $\tau \leq$ ITMAX (say 12) the current time step has to be rejected. A reduced time step size is then computed by

$$\Delta t_n^{reduced} = \Delta t_n / TDIV \tag{1-52}$$

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

1.6.3 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 $\left\|d^{n+1}\right\|_{L_p}$ which is a function of $(X^{n+1} - X_p^{n+1})$, cf. eqns (1-42a), (1-43) and (1-44). The advantage of the PCOSN is that it controls both the truncation and the iteration errors by only one user-

specified tolerance δ. To make the TBFN comparable to the PCOSN scheme we use an equivalent deviatory error norm $\left\|d_\tau^{n+1}\right\|_{L_p}$ as a function of $(X_{\tau+1}^{n+1} - X_\tau^{n+1})$, cf. eqns (1-49a) and (1-49b). 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 $\left\|d_\tau^{n+1}\right\|_{L_p}$, the residual $\left\|R_\tau^{n+1}\right\|_{L_p}$ 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

$$\left\|R_\tau^{n+1}\right\|_{L_p} < \delta_2 \left\|F^{n+1}\right\|_{L_p} \tag{1-53}$$

where a second tolerance δ_2 is introduced and an appropriate normalization of the residual (here with respect to the external supply 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 δ and δ_2, where δ measures the temporal discretization error and δ_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 (1-49a) and, of course, it is easy to replace (1-49a) by (1-53).

In the present study we do not use the condition (1-53). We shall show that the $\left\|d_\tau^{n+1}\right\|_{L_p}$ error norms are sufficient, at least for the examples considered, to ensure the overall evolution of the nonlinear process

under a small global mass balance error $\left\|\boldsymbol{R}_\tau^{n+1}\right\|_{L_2}$. Additionally, we shall observe $\left\|\boldsymbol{R}_\tau^{n+1}\right\|_{L_2}$ in our examples and give estimates of the RMS-based integral (total) mass balance error $\text{TMBE}(T)$ at the final simulation time T in the form

$$\text{TMBE}(T) = \frac{\int_{t=0}^{T} \left\|\boldsymbol{R}_\tau(t)\right\|_{L_2} dt}{\int_{t=0}^{T} \left\|\boldsymbol{F}(t)\right\|_{L_2} dt} \tag{1-54}$$

Eqn (1-54) 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.

1.7 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.*[42] or Oldenburg and Pruess[34] 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 eqn (1-14), where the nodal basis functions $N_I(x_i) = N_I(\xi, \eta, \zeta)$ are evaluated at the element centroid ($\xi = \eta = \zeta = 0$); ξ, η, and ζ 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 eqn (1-14). The evaluation point $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ is used for Gauss integration of the matrix terms (1-16c) and (1-16d) 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 Ψ_p^{n+1} (or Ψ_τ^{n+1} for the TBFN scheme) a specific flux can be computed at a central position of an element e

$$v_e^{n+1} = -\nabla N_I(0, 0, 0) \cdot [\psi_{pI}^{n+1} + (1 + \chi)e_I] \qquad (1\text{-}55)$$

and, the trajectory of the vector v_e^{n+1} can be easily found. Along the trajectory, in the upstream direction, the upstream position $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ is set at the intersection with the element border (Fig. 1.1).
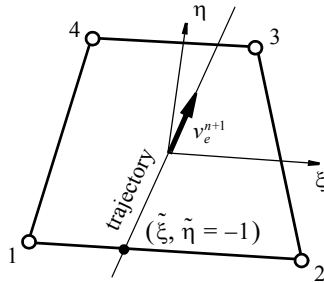
Figure 1.1 Upstream local coordinates $(\tilde{\xi}, \tilde{\eta})$ in a 2D finite element.

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

$$K_r^e(x_i, t) = N_I(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})K_{rI}^e(t) \qquad (1\text{-}56)$$

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 situ-

ation in Fig. 1.1 for a 2D isoparametric finite element, $\tilde{\eta}$ is -1 and K_r^e, from eqn (1-56), becomes independent of nodes 3 and 4, $viz., K_r^e = [(1 - \tilde{\xi})K_{r1}^e + (1 + \tilde{\xi})K_{r2}^e]/2$.

1.8 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 (1-26)[13]

$$\begin{aligned} tol_f &= 0.99 \\ tol_b &= 0.89 \end{aligned} \qquad (1\text{-}57)$$

and the tolerance δ encompassing both the time truncation error measure and the Newton convergence criterion is

$$\delta = 10^{-4} \qquad (1\text{-}58)$$

using the RMS error norm (1-43) 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 δ. 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

$$DSWISH = 0.4$$
$$DPWISH = 4000 \text{ kPa}$$
(1-59)

will be used[13]. In the TBFN solution technique temporal nonlinear discretization errors may occur due to a fast-but-coarse time stepping. The total mass balance errors will be quantified by the $\text{TMBE}(T)$ estimate (1-54).

The large target change parameters (1-59) 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 δ (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.

1.8.1 Infiltration in homogeneous and inhomogeneous soil columns

1.8.1.1 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[41]: column length of 1 m, van Genuchten-Mualem parametric model (1-3), (1-4) in using $n = 2$, ($m = 0.5$), $\alpha = 3.35$ 1/m, $\varepsilon = 0.368$, $s_r = 0.277$, and $s_s = 1.0$, isotropic saturated conductivity of $0.922 \cdot 10^{-4}$ m/s, 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.

Figure 1.2 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 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. 1.2. 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 δ to 10^{-3} the FE/BE scheme still gives identical results, but the AB/TR began to fail in producing nonlinear wiggles.
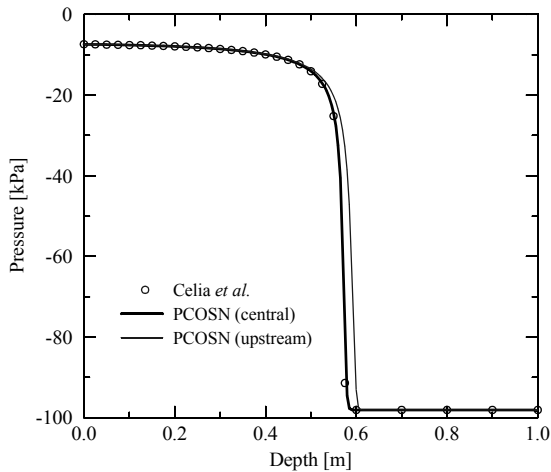
Figure 1.2 Pressure profiles at $t = 1$ day for the dense grid: PCOSN results for central and upstream weighting (both FE/BE and AB/TR scheme) with error $\delta = 10^{-4}$ in comparison with Celia *et al.*'s results[4,41].

Alternatively, if we use a Newton mixed ($\psi - s$)-form scheme, *cf.* eqn (1-30), where the primary variable is always the pressure head ψ, with a FE/BE time marching strategy the same results as outlined in Fig. 1.2 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 1.1 summarizes the solution effort needed for the different predictor-corrector schemes and error tolerances.

Figure 1.3 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. 1.4. The TBFN scheme requires only a small number of Newton steps as summarized in Table 1.2. 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. 1.4). It is important to indicate that this effect is independent of the Newton convergence limit δ. We obtained the same leading curve behavior if decreasing δ (down to 10^{-6}). As given in Table 1.2 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. 1.4). This clearly indicates that the error of the TBFN scheme is caused by temporal discretization, which will be further discussed below.
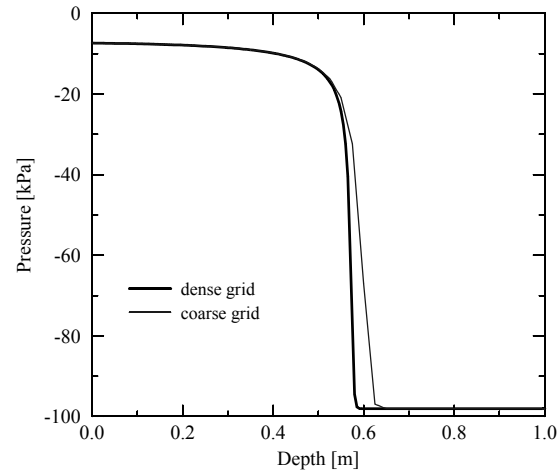
Figure 1.3 Pressure profiles at $t = 1$ day computed by the PCOSN scheme (central weighting) with error $\delta = 10^{-4}$ for the dense and coarse grid.

Figure 1.4 Computed pressure profiles at $t = 1$ day for the PCOSN scheme (with $\delta = 10^{-4}$) and the TBFN scheme (using $\delta = 10^{-4}, ..., 10^{-6}$) at unconstrained ($\Xi_{max} = \infty$) and constrained ($\Xi_{max} = 1.1$) time stepping; dense grid and central weighting.

The time behavior of the residual error $\|R\|_{L_2}$ is plotted in Fig. 1.5 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 $\delta = 10^{-24}$ for a RMS error convergence criterion (1-43). The total mass balance error $\text{TMBE}(T = 1 \text{ d})$, eqn (1-54), can be estimated at $O(10^{-3})$ for the PCOSN and $O(10^{-4})$ for the TBFN.

Table 1.1 Solution effort needed for the PCOSN variable switching scheme compared to the Newton mixed (ψ-*s*-) form solution (dense grid, simulation time 1 day)

Scheme	Type	Weighting	Error δ	Actual time steps	Total Newton steps*)	Efficiency
PCOSN	FE/BE	central	10^{-4}	437	443	1.
PCOSN	FE/BE	upstream	10^{-4}	379	386	.87
PCOSN	FE/BE	central	10^{-3}	283	352	.79
PCOSN	FE/BE	upstream	10^{-3}	148	151	.34
PCOSN	AB/TR	central	10^{-4}	436	580	1.31
PCOSN	AB/TR	upstream	10^{-4}	330	355	.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	.97

*) Including rejected steps

Table 1.2 Solution effort for the TBFN scheme using fully implicit time stepping and central weighting (dense grid, simulation time 1 day)

Error δ	Constraint Ξ_{max}	Weighting	Actual time steps	Total Newton steps*)	Efficiency (Tab. 1.1)
10^{-4}	∞	central	8	88	0.2
10^{-4}	∞	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

*) Including rejected steps
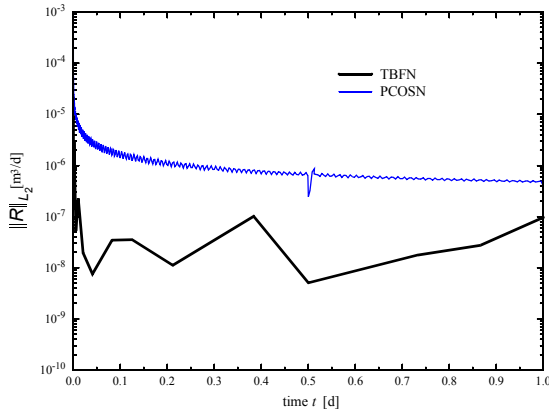
Figure 1.5 History of residual error $\|R\|_{L_2}$ for the TBFN and PCOSN schemes with $\delta = 10^{-4}$, RMS error convergence criterion (1-43) and central weighting.

1.8.1.2   Van Genuchten's problem

Van Genuchten[44] 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 (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 $t \leq 1$ 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 $q_n^{hgrad} = K|_{bottom} = 4$ m/d is imposed[8]. Accordingly, the bottom boundary can freely drain[29]. The parameters of the constitutive relations (van Genuchten-Mualem model) are fully listed in[41].

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. 1.6 for the infiltration period. Differences between central and upstream weighting are also exhibited in Fig. 1.6. 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 ψ^0.

Using the PCOSN scheme with FE/BE and central weighting the computed saturation profiles at $t = 1$ d is shown in Fig. 1.7 for different ψ^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 ψ 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. Figure 1.8 shows the results for both the mixed ($\psi - s$)-form with Newton iteration (comparable to eqn (1-30)) and the standard ψ-form with Picard iteration and chord 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 ψ-form reveals mass-conservative problems (phase lag). The phase lag error dramatically grows at higher initial suctions as evi-

denced in Fig. 1.8 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. 1.8) in comparison to the good PCOSN results (Fig. 1.7). We were not able to find convergent solutions for both the mixed ($\psi - s$)-form and the standard ψ-form at higher suction values ($\psi^0 < -10^3$ m).
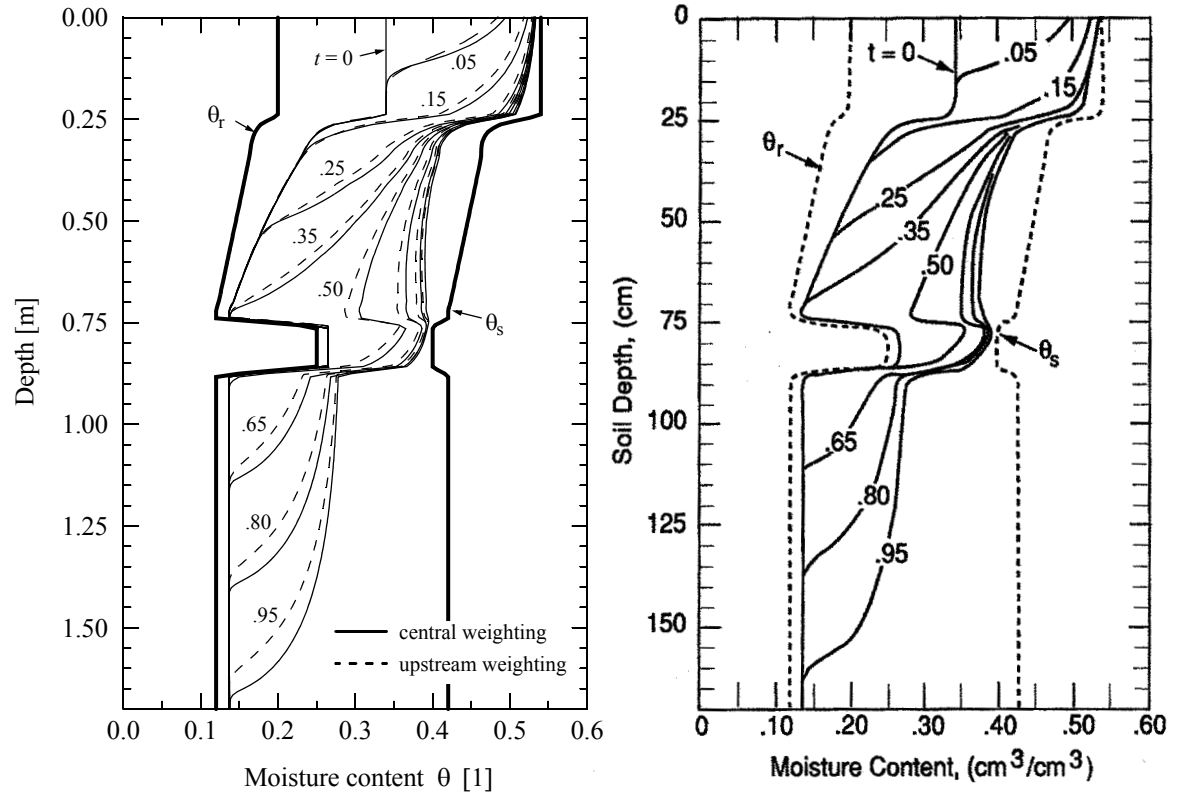
Figure 1.6 Simulated moisture-content profiles ($\theta = s \cdot \varepsilon$) during infiltration: present solutions (left) and van Genuchten's results (right), time in days.

A comparison of the PCOSN and the TBFN variable switching schemes is given in Fig. 1.9. 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 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 1.3 summarizes the solution effort in terms of time steps and number of iterations for different schemes depending on the initial suction ψ^0. The variable switching techniques (PCOSN and TBFN, columns 2-5 of Table 1.3) were successful for all ψ^0 considered, while the schemes using the pressure head ψ as primary variable (mixed Newton ($\psi-s$)-form with both PCOSN and TBFN, and standard Picard ψ-form, columns 6-11 of Table 1.3) have shown unsuitable for very dry conditions $\psi^0 < -10^3$ m. The most interesting outcomes of these comparisons are the following:

Figure 1.8 Saturation distribution at $t = 1$ day computed by the Newton mixed ($\psi-s$)-form and the standard Picard iteration ψ-form (FE/BE, central weighting) with error $\delta = 10^{-4}$ for various initial pressure heads ψ^0 in [m].

Figure 1.7 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 ψ^0 in [m].
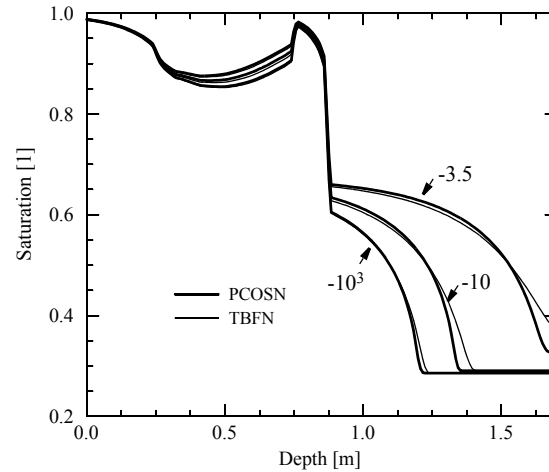
Figure 1.9 Comparison of the PCOSN and the TBFN saturation distributions simulated at $t = 1$ day (FE/BE, central weighting) with error $\delta = 10^{-4}$ for various initial pressure heads ψ^0 in [m].

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. 1.9). 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 ψ^0. At the same time, the number of rejected steps increases so that the overall effort grows with decreasing ψ^0.

The power of the variable switching technique becomes obvious if comparing it with the ψ primary variable solution under the same time stepping strategy.

We additionally applied the TBFN technique to the ψ primary variable form, omitting the variable switching. The computational effort dramatically increases by orders of magnitude (3 to 168 times slower than the TBFN with variable switching as indicated by columns 9 *vs.* 5 of Table 1.3). 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 ψ primary variable form (with the targets (1-59)). Here, the PCOSN scheme is comparable or even faster (*cf.* columns 7 *vs.* 9 in Table 1.3). However, the TBFN scheme was able to find convergent solutions for all ψ^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 $\text{TMBE}(T = 1 \text{ d})$. At lower suction heads ψ^0, see Tab. 1.3, $\text{TMBE}(T = 1 \text{ d})$ is of $O(10^{-4})$ for the PCOSN and $O(10^{-5})$ for the TBFN. At higher suction heads ψ^0 we found $\text{TMBE}(T = 1 \text{ d})$ of $O(10^{-3})$ for the PCOSN and $O(10^{-4})$ for the TBFN.

Table 1.3 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 ψ^0 [m]	Variable switching				Primary variable ψ					
	PCOSN		TBFN[1]		Mixed ($\psi - s$)-form, Newton, eqn (1-30)				Standard ψ-form, Picard, eqn (1-32)	
					PCOSN		TBFN[1]			
	Time steps	Total Newton steps [2]	Time steps	Total Newton steps [2]	Time steps	Total Newton steps [2]	Time steps	Total Newton steps [2]	Time steps	Total Picard steps [2]
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	

[1] Additional time constraint $\Delta t_{max} = 0.05$ d
[2] Including rejected steps

1.8.2 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 α in eqn (1-3) or eqn (1-5)) 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. 1.10. Due to the large α-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[8]). The problem is solved by using both the van Genuchten-Mualem (1-3) and the Brooks-Corey (1-5) constitutive

relationships. The latter offers the advantage to choose the $K_r(s)$ relationship (1-6) independently of the $\psi(s)$-curve (1-5).

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

$$\frac{dh}{dt} = -\frac{K}{\varepsilon(s_s - s_r)}\left[\frac{h(t)}{h(t) + L}\right] \tag{1-60}$$

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

$K = 10^{-4}$ m/s
$\varepsilon = 0.33333$
$S_o = 10^{-12}$ 1/m
$s_s = 1$
$s_r = 0.1$
$n = 2$
$\alpha = 10^4$ 1/m
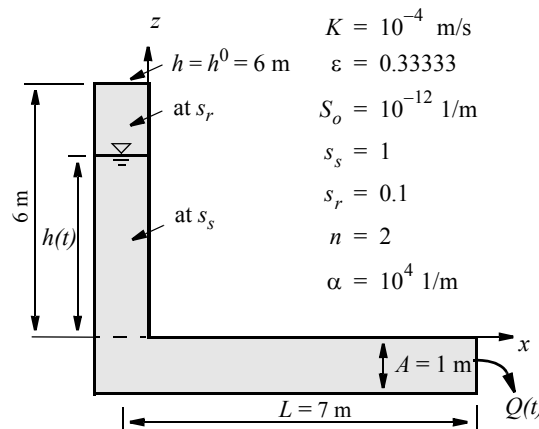
Figure 1.10 Sketch of the drainage problem.

Integrating this equation yields

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

and

$$Q(t) = KA\left(\frac{\varphi}{\varphi + L/h^0}\right) \tag{1-61b}$$

Table 1.4 lists the analytical results at given relative drawdown φ. The domain is discretized in 200 quadrilateral finite elements ($\Delta z = 6.5$ cm), where the original problem (Fig. 1.10) 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.

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. 1.10 only the variable switching technique was successful while the mixed ($\psi - s$)-form ran into significant convergence difficulties and the standard ψ-form even completely failed. The computational results for the PCOSN scheme are listed in Table 1.5. The agreement with the analytical results (Table 1.4) 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 (1-5) and (1-6) 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 κ to unity is somewhat artificial. However, it is acceptable for this water table problem (see the results presented in Table 1.5 in comparison to the analytical results of Table 1.4).

Table 1.4 Analytical results

φ	t [d]	Q [m³/d]	$\int_0^t Q(t)dt$
1.	0	3.987692	0
.75	.122006	3.380870	.45
.50	.272640	2.592000	.9
.25	.493197	1.524706	1.35
.0372872	1.	.267586	1.789
0	∞	0	1.8

Table 1.5 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 [m³/d]	$\int_0^t Q(t)dt$	Q [m³/d]	$\int_0^t Q(t)dt$	Q [m³/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	$4.0 \cdot 10^{-8}$
.122006	3.3669	.4454	3.2917	.4407	3.2715	.4394
.272640	2.6185	.8884	2.5026	.8783	2.4722	.8722
.493197	1.5803	1.328	1.4703	1.313	1.4434	1.300
1.	.3285	1.727	.2742	1.686	.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 1.5. The number of Newton steps slightly increased to 2818 for a 1-day simulation if upstream weighting was applied.

In estimating the $\text{TMBE}(T = 1\ \text{d})$ error (1-54) 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 1.5 to the analytical results of Table 1.4.

1.8.3   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. 1.11. Water infiltrates with a very large rate into a dry soil at ψ^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 1.6 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 50x60 quadrilateral mesh (3111 nodes) according to the spatial discretization used by Kirkland et al.[26] and Forsyth et al.[13].
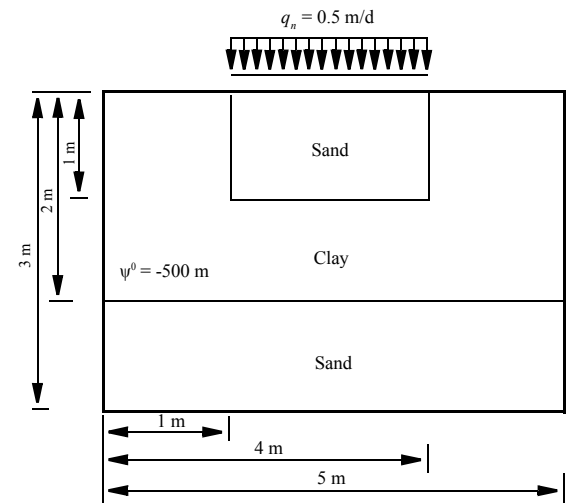
Figure 1.11 Perched water table problem (modified from[26]).

Table 1.6 Material properties for the perched water table problem

Material	K [m/s]	ε [1]	s_r [1]	α [1/m]	n [1]
Sand	$6.262 \cdot 10^{-5}$	.3658	.07818	2.80	2.2390
Clay	$1.516 \cdot 10^{-6}$	.4686	.2262	1.04	1.3954

A comparison of the pressure contours at 1 day with Kirkland *et al.*'s results reveals an acceptable agreement as displayed in Fig. 1.12. 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. 1.13). Forsyth *et al.*[13] used an aggressive target-based time marching scheme similar to 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. 1.12 and 1.13, respectively.

As displayed in the time step histories for both schemes in Fig. 1.14 the TBFN scheme progresses faster at the beginning, while the PCOSN scheme takes smaller step sizes due to the temporal discretization 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.
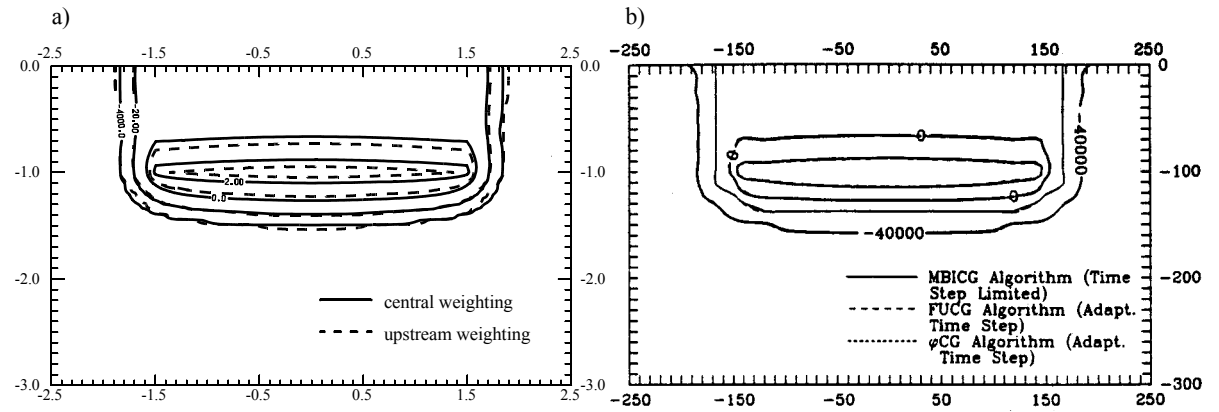
Figure 1.12 Simulated pressure contours at $t = 1$ d: a) present results, PCOSN and TBFN, FE/BE, central and upstream weighting, pressure contours in [kPa], lengths in [m]; b) Kirkland *et al.*'s results[26], pressure head contours in [cm], lengths in [cm].
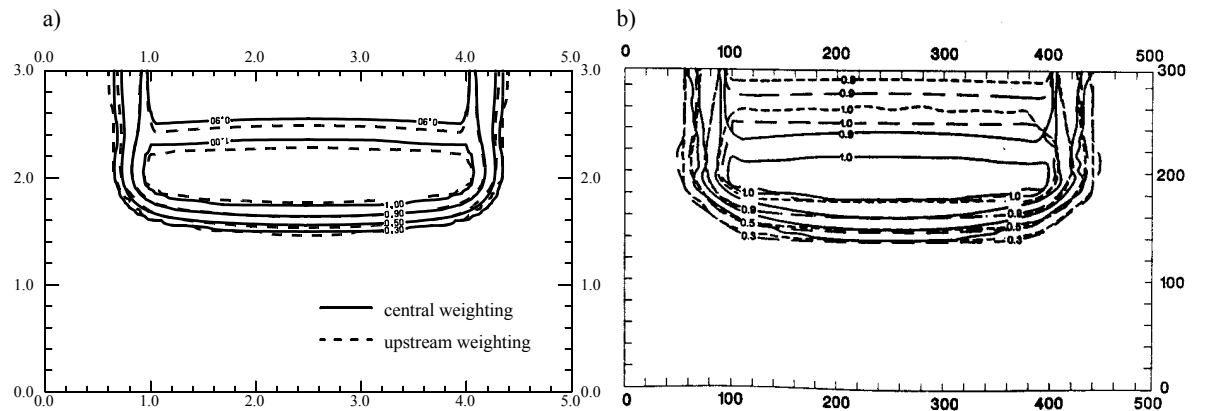
Figure 1.13 Simulated saturation contours at $t = 1$ d: a) present results, PCOSN and TBFN, FE/BE, central and upstream weighting, lengths in [m]; b) Forsyth *et al.*'s results[13]; - - - one phase, upstream weighting; . . . one phase, central weighting; _____ two phases, upstream weighting, lengths in [cm].
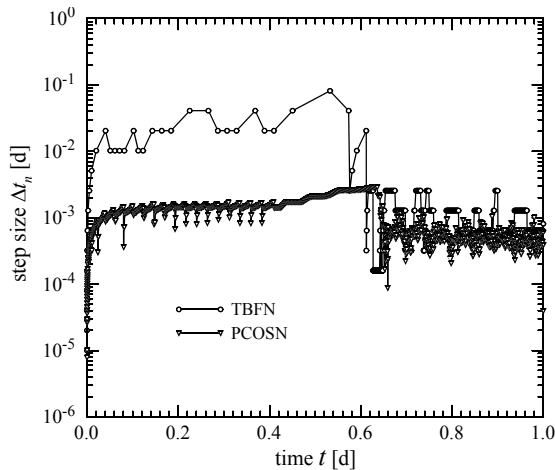
Figure 1.14 Time step histories of the perched water table problem for the TBFN and PCOSN schemes (FE/BE, central weighting) using $\delta = 10^{-4}$ and $\Delta t_0 = 10^{-5}$ d ($\Xi_{max} = 2$ for TBFN); required time steps: 582 (Newton 3381) for TBFN and 1211 (Newton 1556) for PCOSN.

The simulations with the PCOSN and TBFN schemes give identical results (Figs. 1.12 and 1.13) 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 (1-54) was found to be of $O(10^{-4})$ for the PCOSN and of $O(10^{-5})$ for the TBFN scheme.

1.8.4   Infiltration in a large caisson

1.8.4.1   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. Figure 1.15 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 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 90x21 quadrilateral elements (1890 nodes) as in Forsyth *et al.*[13].

Based on the given control parameters the TBFN scheme was about four times faster than the PCOSN scheme as indicated in Table 1.8. On the average 3 to 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. 1.16 and 1.17 for $\psi^0 = -7.34$ m and $\psi^0 = -100$ m, respectively, in comparison to Forsyth *et al.*'s findings[13].

Table 1.7 Material properties for Forsyth *et al.*'s problem (van Genuchten-Mualem parametric model)

Zone	K [m/s]	ε [1]	s_r [1]	α [1/m]	n [1]
1	$9.153 \cdot 10^{-5}$	.3680	.2771	3.34	1.982
2	$5.445 \cdot 10^{-5}$	.3510	.2806	3.63	1.632
3	$4.805 \cdot 10^{-5}$	.3250	.2643	3.45	1.573
4	$4.805 \cdot 10^{-4}$	.3250	.2643	3.45	1.573

Surprisingly, the PCOSN results are rather depart from the TBFN results, especially for the case ψ^0 = -7.34 m. The saturation front is significantly diffused by the 'low-cost' TBFN simulation using the 'aggressive' control parameters (1-59) 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 ψ^0 = -7.34 m and 48 steps at ψ^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 (1-59) have to be chosen for the TBFN to enforce smaller time step sizes and to find results comparable to the PCOSN.
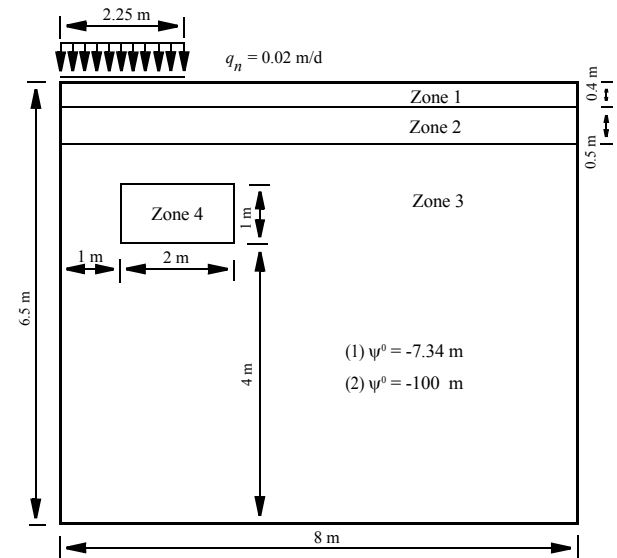
Figure 1.15 Forsyth *et al.*'s infiltration problem (modified from[13]).

Table 1.8 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
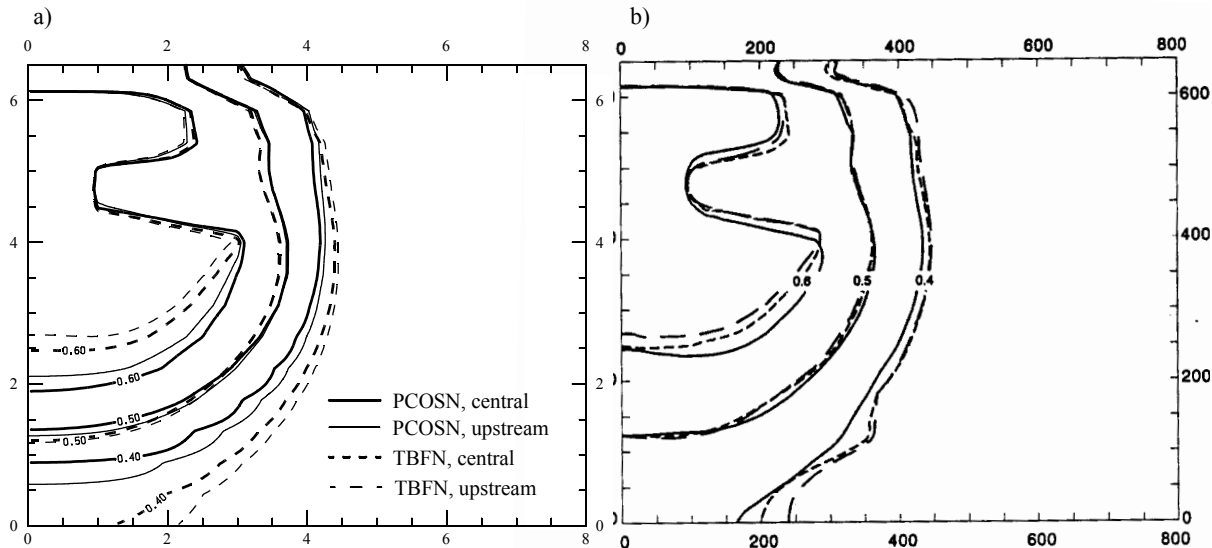
Figure 1.16 Computed saturation contours at t = 30 d, initial pressure head ψ^0 = -7.34 m: a) present solutions by PCOSN and TBFN, central and upstream weighting, lengths in [m]; b) Forsyth *et al.*'s results[13]; - - - one phase, upstream weighting; . . . one phase, central weighting; _____ two phases, upstream weighting, lengths in [cm].

It is apparent that the present problem is sensitive to discretization errors. The influence of the spatial discretization is illustrated in Fig. 1.18 for the case $\psi^0 = -100$ m. The results of structured coarse meshes (90x21 and 21x90 nodes) are compared to a dense unstructured mesh consisting of 56,960 triangular elements (28,917 nodes). This dense mesh is generated by splitting each quadrilateral into two triangles followed by a double total refinement into four triangles (20x89x2x4x4). It shows how a coarse meshing in one direction can lead to phase lag errors and smearing of the saturation profiles.
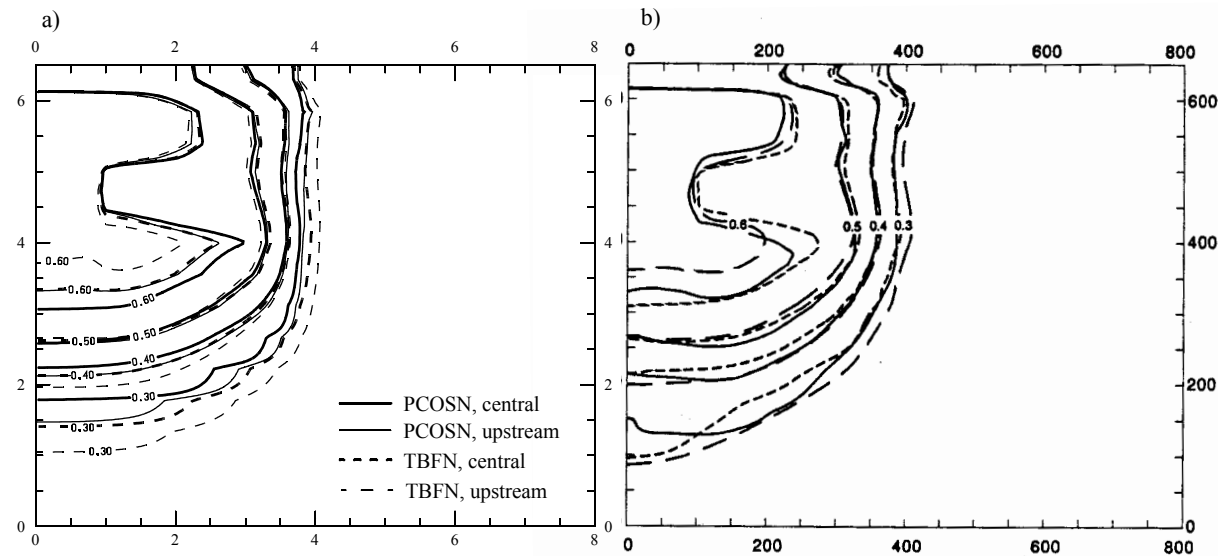
Figure 1.17 Computed saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -100$ m: a) present solutions by PCOSN and TBFN, central and upstream weighting, lengths in [m]; b) Forsyth *et al.*'s results[13]; - - - one phase, upstream weighting; . . . one phase, central weighting; ____ two phases, upstream weighting, lengths in [cm].
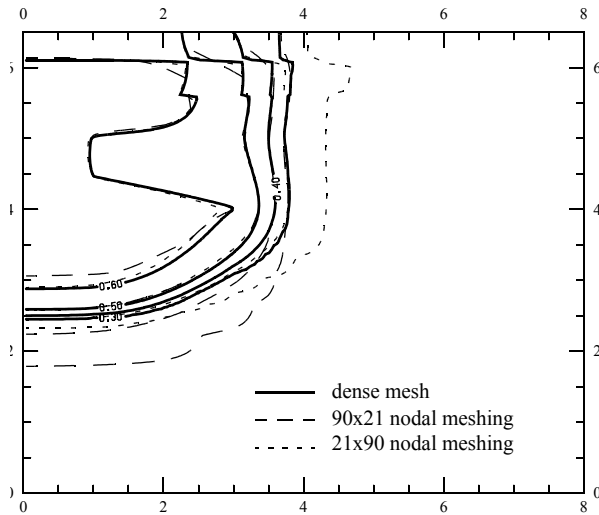
Figure 1.18 Influence of spatial discretization, computed saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -100$ m: dense mesh consists of 56,960 triangles and 28,917 nodes, central weighting, PCOSN (FE/BE) scheme with 2507 implicit time steps and 3596 Newton steps, lengths in [m].

The history of the residual error $\|R\|_{L_2}$ depending on the selected time stepping schemes and the initial pressure head ψ^0 is plotted in Fig. 1.28. 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.
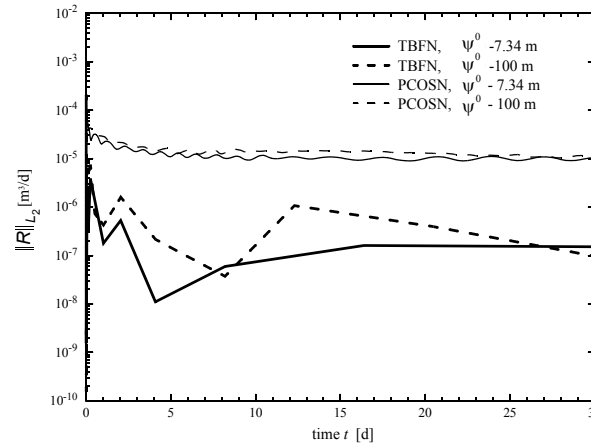
Figure 1.19 History of residual error $\|R\|_{L_2}$ for the TBFN and PCOSN schemes with $\delta = 10^{-4}$, RMS error convergence criterion (1-43), central weighting and 90x21 mesh.

1.8.4.2 Forsyth and Kropinski's problem

Forsyth and Kropinski[14] modified the above infiltration problem of Fig. 1.15 by increasing the pore size distribution index n to 5 for the zones 3 and 4. The other parameters remain unchanged and correspond to Table . 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. 1.20 for a structured 90x21 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. 1.20 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. Figure 1.21 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. 1.20). Differences in the lag of the saturation pro-

file 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., 21x90 instead of 90x21) can considerably improve the results (Figs. 1.22(a) and 1.22(b)). The solution can be compared to the results obtained with the dense mesh (28.917 nodes) shown in Fig. 1.23. 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. 1.22. In contrast, Forsyth and Kropinski[14] predict a lead in the saturation pattern (Fig. 1.22(b)).
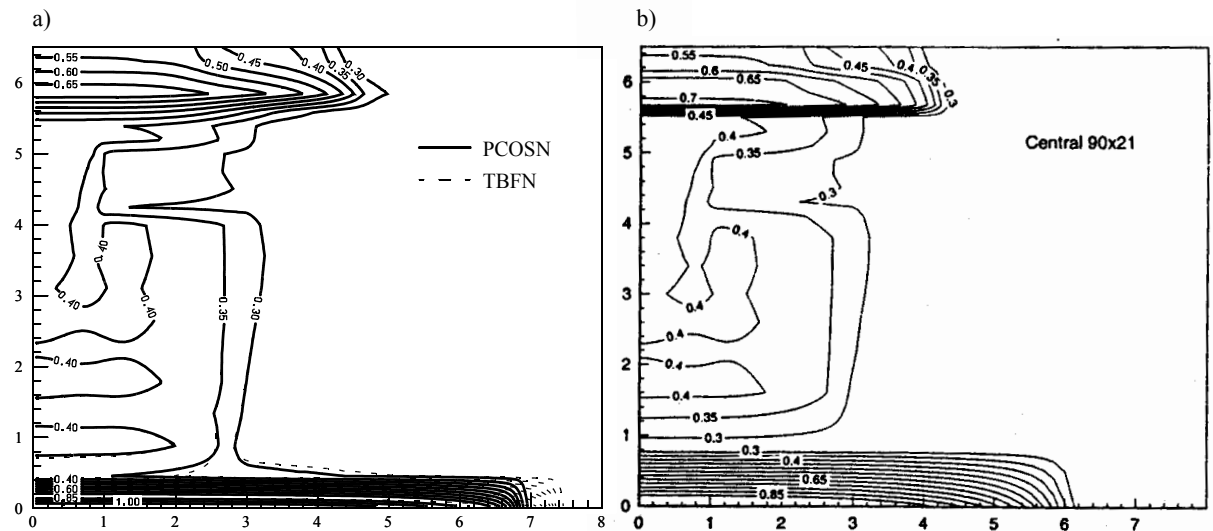
Figure 1.20 Saturation contours at $t = 30$ d, initial pressure head $\psi^0 = -100$ m, and central weighting: a) present solutions by PCOSN and TBFN, 90x21 nodal meshing; b) Forsyth and Kropinski's results[14]; lengths in [m].

a)

b)
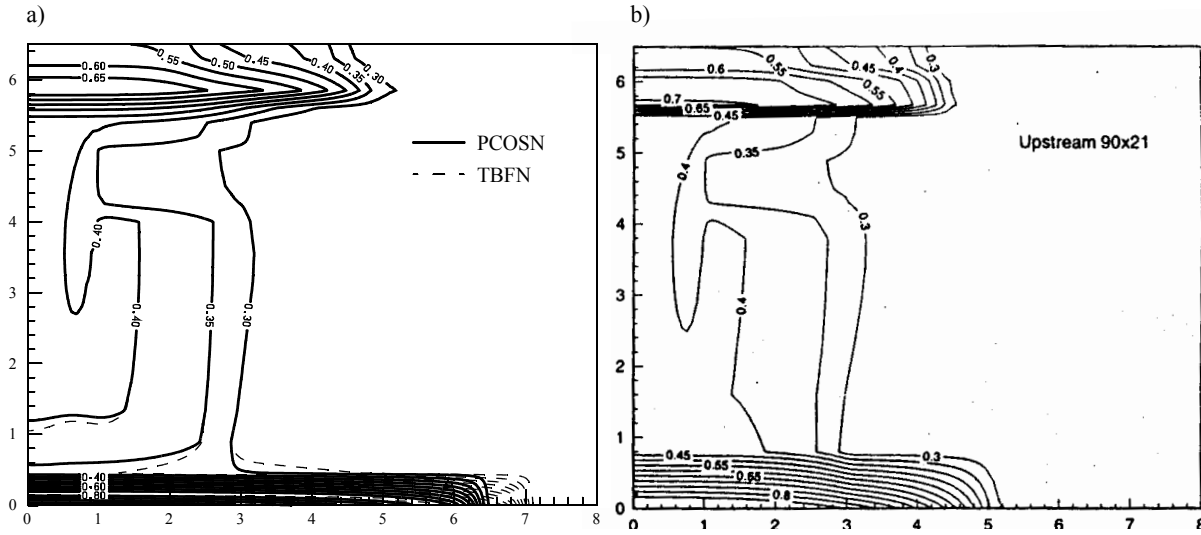
Figure 1.21 Saturation contours at t = 30 d, initial pressure head ψ^0 = -100 m, and upstream weighting: a) present solutions by PCOSN and TBFN, 90x21 nodal meshing; b) Forsyth and Kropinski's results[14]; lengths in [m].
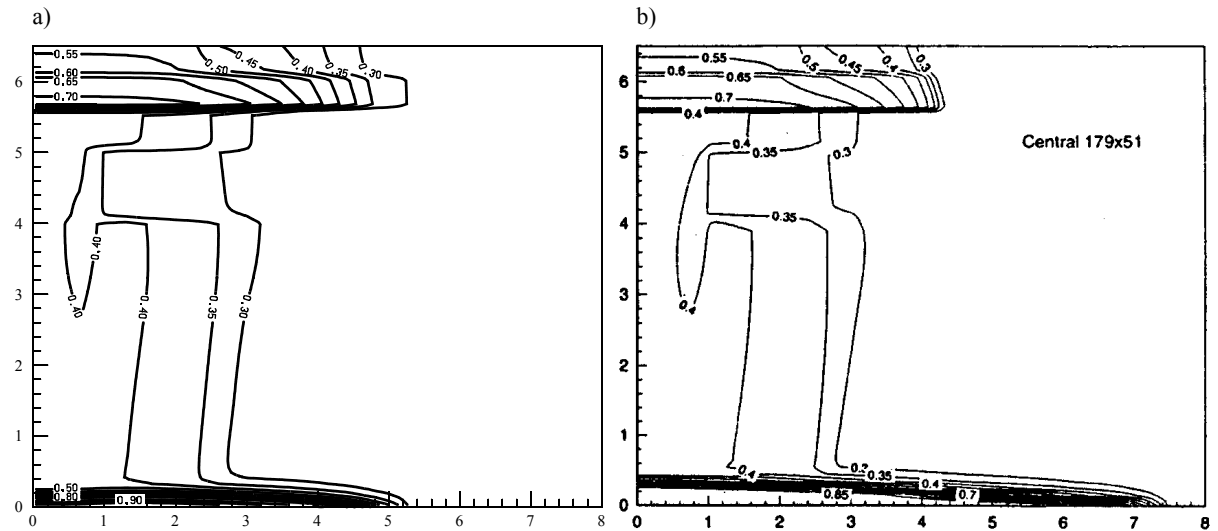
Figure 1.22 Saturation contours for refined meshes at $t = 30$ d, initial pressure head $\psi^0 = -100$ m, and central weighting: a) present solutions by PCOSN, 21x90 nodal meshing; b) Forsyth and Kropinski's results[14]; lengths in [m].

In checking the mass balance errors $\mathrm{TMBE}(T)$, eqn (1-54), we estimate the same order as indicated in the above problem of section 1.8.4.1: $\mathrm{TMBE}(T = 30\ \mathrm{d})$ of $O(10^{-5})$ for the PCOSN and of $O(10^{-6})$ for the TBFN.
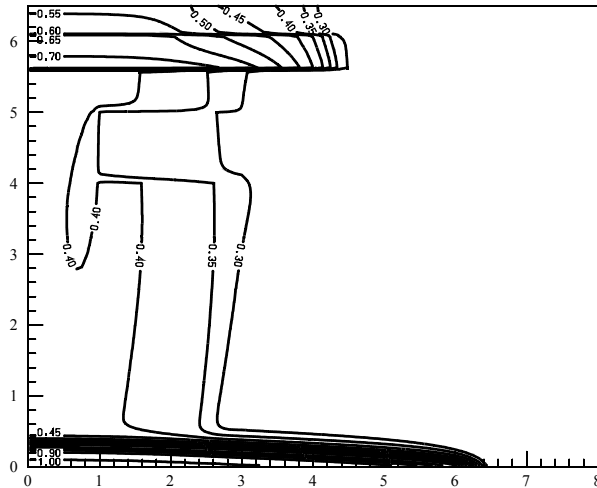
Figure 1.23 Present saturation contours for the dense mesh (28,917 nodes) at t = 30 d, initial pressure head ψ^0 = -100 m, central weighting, lengths in [m].

1.8.5   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[34,48]. 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 of the capillary diversion, the TBFN scheme seems be the most effective solution technique for this class of problems.

1.8.5.1   Webb's problem

Oldenburg and Pruess[34] 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[40] could be achieved. More recently, Webb[48] could improve the steady-state results by using an upstream weighting technique agreeing well with Ross' analytical prediction[40]. We use Webb's capillary barrier problem[48] 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 1.9. 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 1.9) we found a better convergence behavior for a potential-type boundary condi-

tions, where the hydraulic head h is imposed. Since the α-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 1.9). The model domain is appropriately discretized in quadrilateral elements as displayed in Fig. 1.24. 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.

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 1.8.3 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 (1-44) instead of the integral RMS norm (1-43). Here, an error tolerance of $\delta = 10^{-3}$ turned out to be sufficient.

Table 1.9 Material properties for Webb's capillary barrier problem (van Genuchten-Mualem parametric model)

Parameter	Upper layer (fine)	Lower layer (coarse)
ε [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.	1.
n [1]	5.74	2.19
α [1/m]	3.9	490.
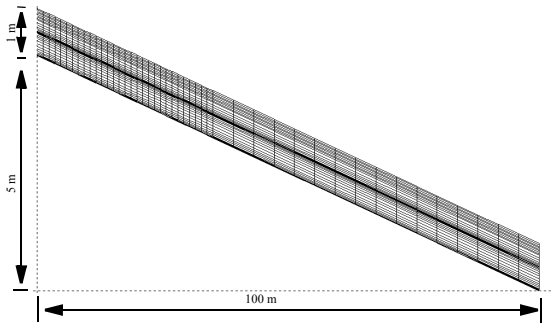
Figure 1.24 Model domain and mesh (1472 quadrilaterals with 1551 nodes) for Webb's capillary barrier problem[48] (exaggeration 10 : 1).

Figure 1.25 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. 1.25. 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[40] and the numerical results obtained by Webb[48] 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[48] computed 33.2 m). As evidenced in Fig. 1.26 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. 1.26 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 classic ψ-based form becomes an effective alternative. We confirmed the above solutions for the ψ-model (1-32), using the predictor-corrector time stepping scheme for both FE/BE and AB/TR, and without the Newton method.
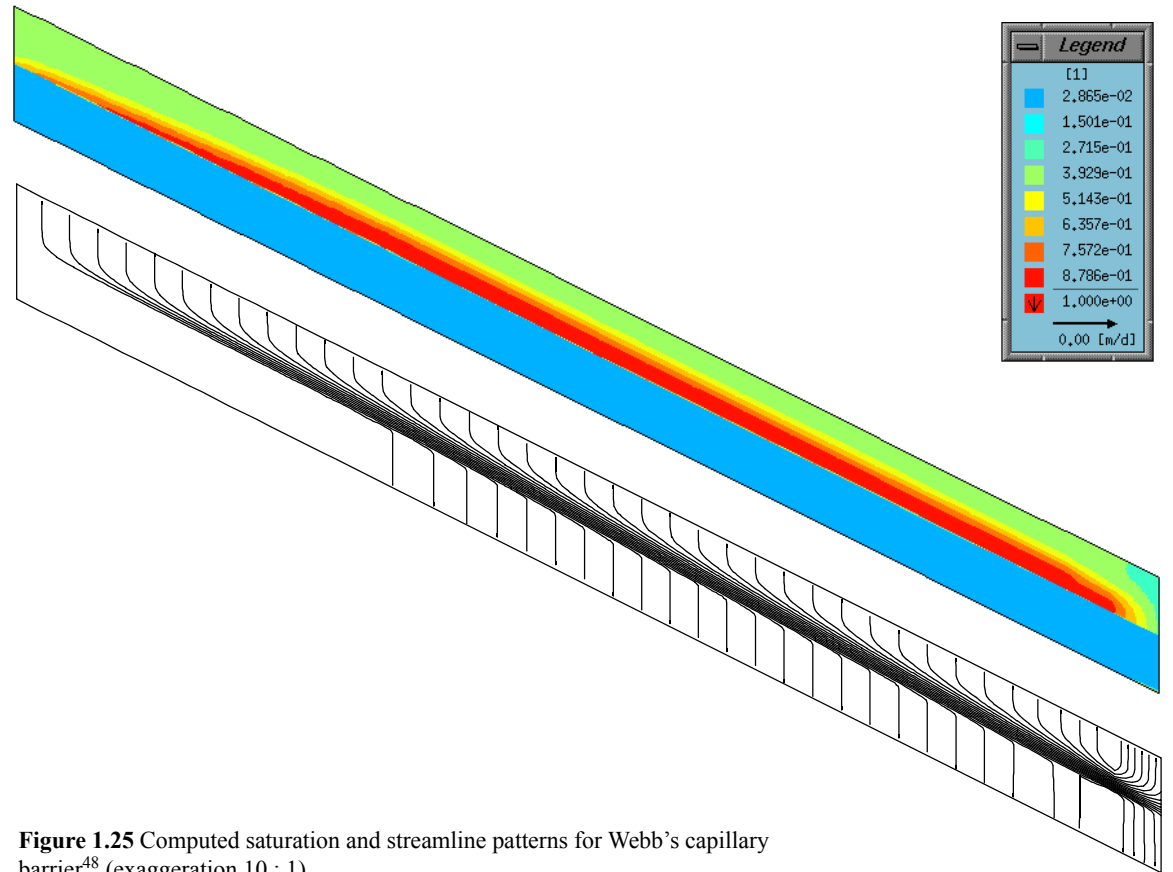
Figure 1.25 Computed saturation and streamline patterns for Webb's capillary barrier[48] (exaggeration 10 : 1).
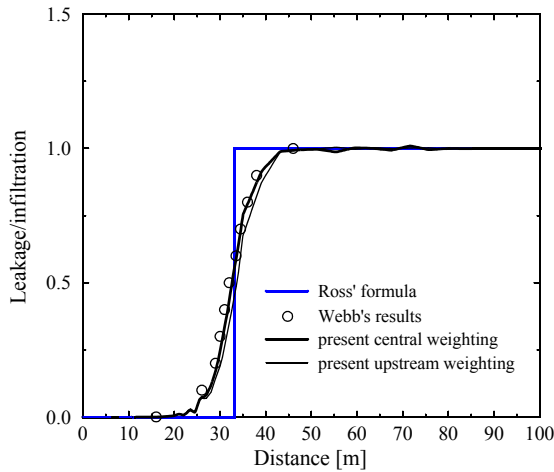
Figure 1.26 Leakage/infiltration ratio in the coarse layer for both central and upstream weighting compared to Ross' analytical formula[40] and Webb's numerical results[48].

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 $q_n^{hgrad} = K|_{bottom} = 0.23985$ m/d applies there.

1.8.5.2 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. 1.27. The material properties and the initial pressure conditions for the different layers are given in Table . 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. 1.27. 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
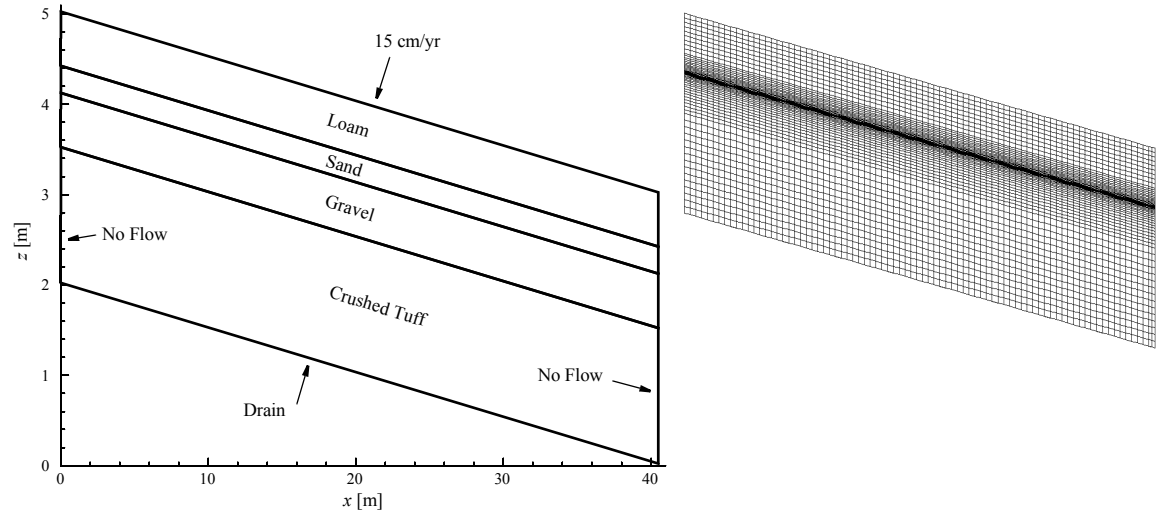
Figure 1.27 Capillary barrier model domain (modified from[14]) and mesh (5002 isoparametric bilinear elements with 5146 nodes).

Table 1.10 Material properties and initial pressure for Forsyth and Kropinski's capillary barrier problem (van Genuchten-Mualem parametric model)

Zone	K [m/s]	ε [1]	s_r [1]	α [1/m]	n [1]	ψ^0 [kPa]
Loam	$1.668 \cdot 10^{-5}$	.452	.0752	4.3	1.246	-10^6
Sand	$6.573 \cdot 10^{-5}$	.345	.046	6.34	1.53	-10^6
Gravel	$3.502 \cdot 10^{-3}$	.419	.074	469.	2.57	-30
Crushed tuff	$2.776 \cdot 10^{-6}$	.345	.032	1.43	1.506	$-6 \cdot 10^{10}$

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 years with the TBFN and applying both the L_2 (1-43) and L_∞ (1-44) error norms for terminating the Newton iteration. The evolution of the residual error $\|R\|_{L_2}$ for both norms is depicted in Fig. 1.28. It reveals that the L_2 criterion produces residuals in the range 10^{-3} to 10^{-2} [m^3/d]. For this case the integral total mass balance error is $\mathrm{TMBE}(T = 30\ \mathrm{yr}) \approx 1.2 \cdot 10^{-2}$, which cannot be tolerated. The results for the L_∞ criterion is better by about one order (*cf.* Fig. 1.28) and gives $\mathrm{TMBE}(T = 30\ \mathrm{yr}) \approx 4.7 \cdot 10^{-3}$. Accordingly, only the results obtained under the L_∞ criterion will be discussed.
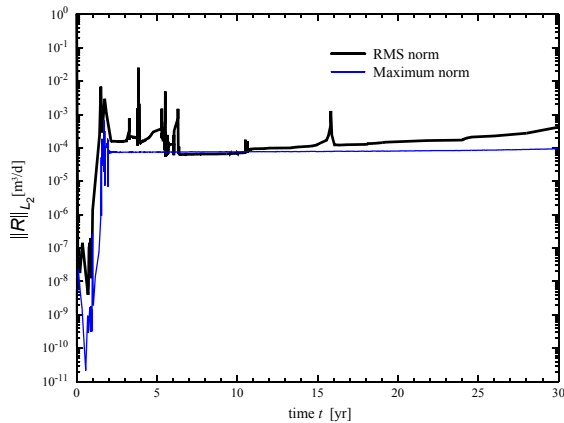
Figure 1.28 History of residual error $\|R\|_{L_2}$ for the RMS L_2 and maximum L_∞ convergence criteria with $\delta = 10^{-4}$ and central weighting.

The 30-year simulation under the L_∞ 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 Figs. 1.29(a), 1.29(a) and 1.31.

Forsyth and Kropinski[14] used both central and upstream weighting at two grid resolutions (52x46 and 103x92). They predict that the capillary barrier fails with a diversion length of about 10 m characterized by a saturation distribution as exemplified in Fig. 1.29(b) for upstream weighting and Fig. 1.29(b) for central weighting with the 52x46 grid.

The present simulations confirm Forsyth and Kropinski's results[14]. The computed saturation distributions are displayed for three specific contour levels in Fig. 1.29(a) for the upstream weighting and in Fig. 1.29(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. 1.27. For the central weighting (Fig. 1.29(a)) we note a jagged saturation profile which disappears for upstream weighting (Fig. 1.29(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. 1.29(b)).
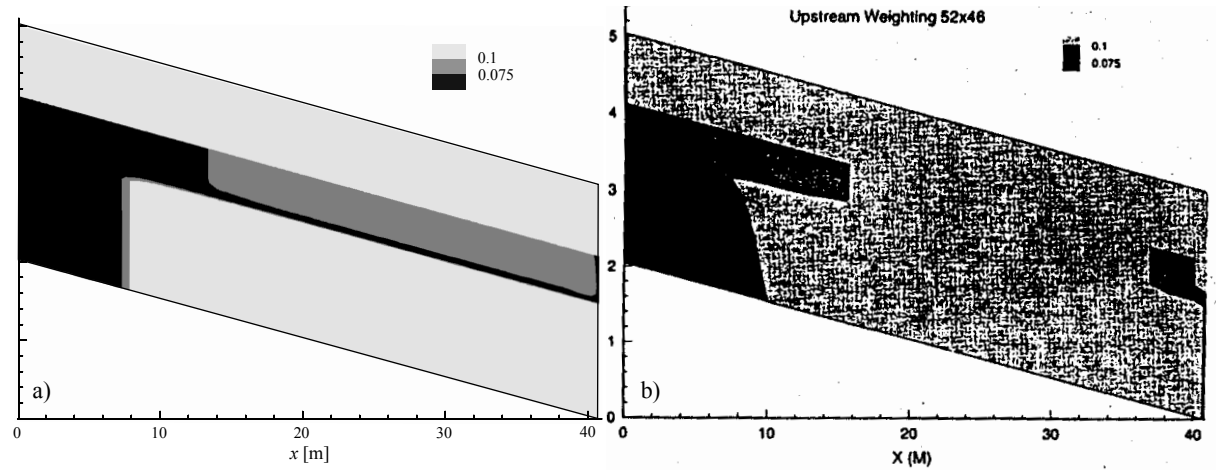
Figure 1.29 Simulated saturation patterns at $t = 30$ yr: a) present solution by TBFN and upstream weighting, b) upstream weighting solution obtained by Forsyth and Kropinski[14].
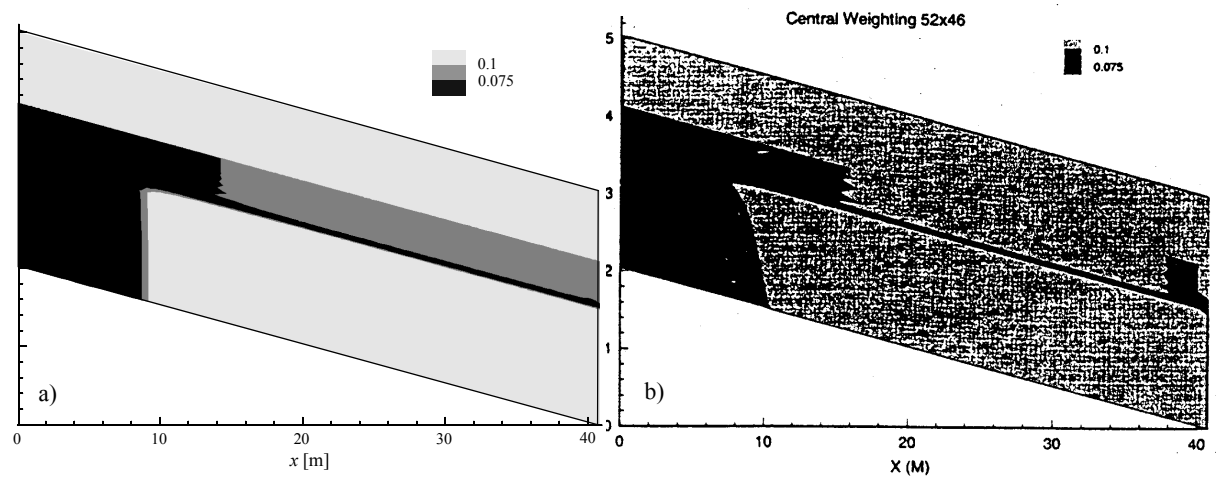
Figure 1.30 Simulated saturation patterns at $t = 30$ yr: a) present solution by TBFN and central weighting, b) central weighting solution obtained by Forsyth and Kropinski[14].

The streamline patterns in Fig. 1.31 illustrate the effect of the capillary barrier at the sand-gravel 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. 1.32. 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[40] using the above van Genuchten parameter for the sand and gravel zones (Table) gives a diversion length of 17.9 m. This value is in good agreement with the present numerical simulations as seen in Fig. 1.32.
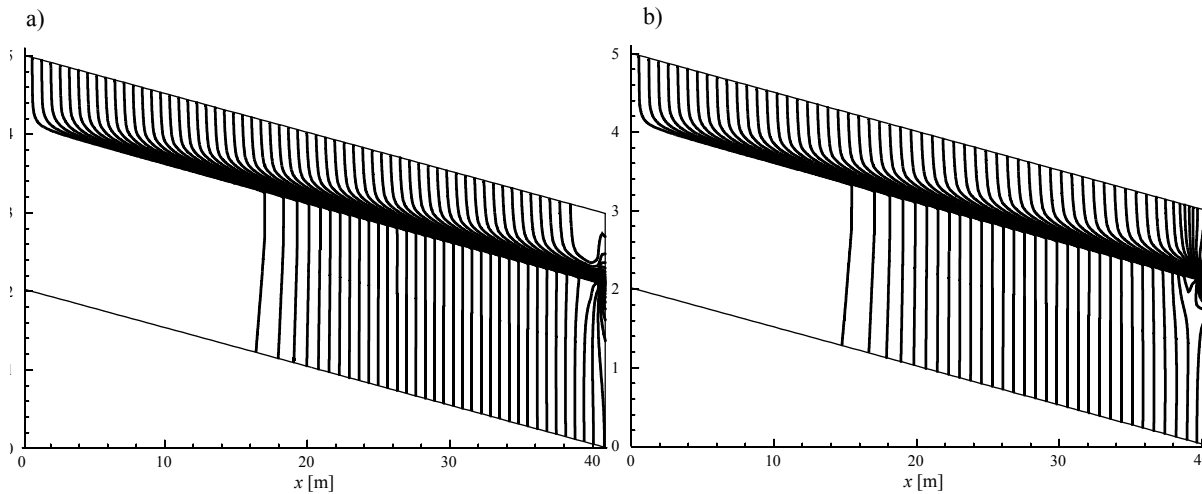
Figure 1.31 Simulated streamline patterns at $t = 30$ yr, TBFN for a) central and b) upstream weighting.
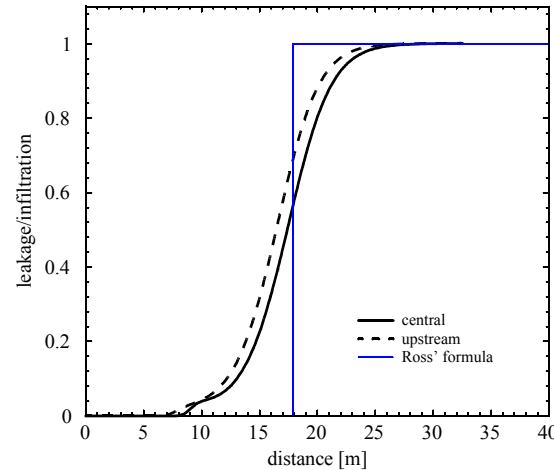
Figure 1.32 Leakage/infiltration ratio in the tuff layer.

1.9 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 ψ-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 tech-

nique 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[12,13,14,24] 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 δ, 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 1.8.1.2).

(2) The primary variable switching technique provides a much better convergence behavior compared to both the mixed $\psi - s$-form and the standard ψ-form of the Richards equation. This is independent of the used time marching scheme (*cf.* Table 1.3). The efficiency of the primary variable switching technique grows with decreasing initial pressure ψ^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 1.8.1.1, 1.8.1.2, 1.8.4.1, and 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 1.8.2 and 1.8.3), 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. 1.16, 1.17 and 1.18; Figs. 1.3, 1.4 and 1.9). 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 ψ-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 $\dot{X}$ occurring in both the Jacobian and the residual (see eqns (1-20), (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. 1.20 and 1.21). 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 1.8.5.1 and 1.8.5.2).

(9) The deviatory convergence criteria in form of L_2 (1-43) and L_∞ (1-44) 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_∞ 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 eqn (1-53) 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

References

1. Allen, M. B., & Murphy, C. A finite element collocation method for variably saturated flows in porous media. *Numer. Methods Partial Differential Equations*, 1985, **3** 229-239.
2. Bear, J., & Bachmat, Y. Introduction to modeling of transport phenomena in porous media, Kluwer Academic Publ., Dordrecht, 1991.

3. Bixler, N. E. An improved time integrator for finite element analysis. *Communications Appl. Num. Meths.*, 1989, **5** 69-78.

4. Celia, M. A., Bouloutas, E. T. & Zarba, R. L. A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, 1990, **26**(7) 1483-1496.

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.-J. G. & Kolditz, O. On finite-element analysis of spatio-temporal buoyancy-driven convection processes in porous media. *Calibration and Reliability in Groundwater Modelling*, Proc. ModelCARE 96 Conf., IAHS Publ. no. 237, 1996, 407-415.

7. Diersch, H.-J, G. A shock-capturing finite-element technique for unsaturated-saturated flow and transport problems, in: V. N. Burganos *et al.* (Eds.), *Computational Methods in Water Resources XII*, vol. 1, Computational Mechanics Publications, Southampton, 1998, 207-214.

8. Diersch, H.-J. G. Treatment of free surfaces in 2D and 3D groundwater modeling. *Mathematische Geologie*, 1998, **2** 17-43.

9. Diersch, H.-J. G. Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.

10. Diersch, H.-J. G. & Kolditz, O. Coupled groundwater flow and transport: 2. Thermohaline and 3D convection systems. *Adv. Water Resour.*, 1998, **21** 401-425.

11. Engelman, M. S., Strang, G. & Bathe, K.-J. The application of quasi-Newton methods in fluid mechanics. *Intern. J. Num. Meths. Engng.*, 1981, **17** 707-718.

12. Forsyth, P. A. & Simpson, R. B. A two-phase, two-component model for natural convection in a porous medium. *Intern. J. Num. Meths. Fluids*, 1991, **12** 655-682.

13. Forsyth, P. A., Wu, Y. S. & 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, P. A. & Kropinski, M. C. Monotonicity considerations for saturated-unsaturated subsurface flow. *SIAM J. Sci. Comput.*, 1997, **18**(5) 1328-1354.

15. Frind, E. O. & Verge, M. Three-dimensional modeling of groundwater flow systems. *Water Resour. Res.*, 1978, **14**(5) 844-856.

16. Fuhrmann, J. On numerical solution methods for nonlinear parabolic problems. *Notes on Numerical Fluid Mechanics*, vol. 59, Vieweg, 1997, 170-180.

17. Gresho, P. M., Lee, R. L. & Sani, R. L. On the time-dependent solution of the incompressible Navier-Stokes equations in two and three dimensions. *Preprint UCRL-83282*, Lawrence Livermore Lab., Univ. Calif., 1979.

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

19. Hillel, D. Fundamentals of soil physics, Academic, San Diego, Calif., 1980.

20. Hills, R. G., Hudson, D. B., Porro, I. & Wierenga, P. J. Modeling one-dimensional infiltration into very dry soils, 1, Model development and evaluation. *Water Resour. Res.*, 1989, **25**(6) 1259-1269.

21. Hughes, T. J. R. A simple scheme for developing 'upwind' finite elements. *Intern. J. Num. Meths. Engng.*, 1978, **12** 1359-1365.

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

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

24. Huyakorn, P. S., Panday, S. & Wu, Y. S. A three-dimensional multiphase flow model for assessing NAPL contamination in porous and fractured media, 1. Formulation. *J. Contaminant Hydrology*, 1994, **16** 109-130.

25. Ju, S.-H. & Kung, K.-J. S. Mass types, element orders and solution schemes for the Richards equation. *Computers & Geosciences*, 1997, **23**(2) 175-187.

26. Kirkland, M. R., Hills, R. G. & Wierenga, P. J. Algorithms for solving Richards' equation for variably saturated soils. *Water Resour. Res.*, 1992, **28**(8) 2049-2058.

27. LaBolle, E. M. & Clausnitzer V. Comment on Russo[1991], Serrano[1990,1998], and other applications of the water-content-based form of Richards' equation to heterogeneous soils. *Water Resour. Res.*, 1999, **35**(2) 605-607.

28. Lehmann, F. & Ackerer, Ph. Comparison of iterative methods for improved solutions of the fluid flow equation in partially satu-

rated porous media. *Transport in Porous Media*, 1998, **31**(3) 275-292.

29. McCord, J. T. Application of second-type boundaries in unsaturated flow modeling. *Water Resour. Res.*, 1991, **27**(12) 3257-3260.

30. Miller, C. T., Williams, G. A., Kelley, C. T. & Tocci, M. D. Robust solution of Richards' equation for nonuniform porous media. *Water Resour. Res.*, 1998, **34**(10) 2599-2610.

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

32. Neuman, S. P. Saturated-unsaturated seepage by finite elements. *J. Hydraul. Div., ASCE*, 1973, **99**(HY12) 2233-2250.

33. 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, 559-566.

34. Oldenburg, C. M. & Pruess, K. On numerical modeling of capillary barriers. *Water Resour. Res.*, 1993, **29**(4) 1045-1056.

35. Panday, S., Forsyth, P. A., Falta, R. W., Wu, Y.-S. & Huyakorn, P. S. Considerations for robust compositional simulations of subsurface nonaqueous phase liquid contamination and remediation. *Water Resour. Res.*, 1995, **31**(5) 1273-1289.

36. Paniconi, C., Aldama, A. A. & Wood, E. F. Numerical evaluation of iterative and noniterative methods for the solution of the nonlinear Richards equation. *Water Resour. Res.*, 1991, **27**(6), 1147-1163.

37. 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) 3357-3374.

38. Rathfelder, K. & Abriola, L. M. Mass conservative numerical solutions of the head-based Richards equation. *Water Resour. Res.*, 1994, **30**(9) 2579-2586.

39. Romano, N., Brunone, B. & Santini, A. Numerical analysis of one-dimensional unsaturated flow in layered soils. *Adv. Water Resour.*, 1998, **21** 315-324.

40. Ross, B. The diversion capacity of capillary barriers. *Water Resour. Res.*, 1990, **26**(10) 2625-2629.

41. Segol, G. Classic groundwater simulations: Proving and improving numerical models, PTR Prentice Hall, Englewood Cliffs, New Jersey, 1994.

42. Simunek, J., Vogel, T. & van Genuchten, M. Th. The SWMS-2D code for simulating water flow and solute transport in two-dimensional variably saturated media. *Research Report No. 126*, U.S. Salinity Lab., Riverside, Calif., 1992.

43. Tocci, M. D., Kelley, C. T. & Miller, C. T. Accurate and economical solution of the pressure-head form of Richards' equation by the method of lines. *Adv. Water Resour.*, 1997, **20** 1-14.

44. Van Genuchten, M. Th. Numerical solutions of the one-dimensional saturated-unsaturated flow equation. Research Rep. No. 78-WR-09, Water Resources Program, Princeton Univ., 1978.

45. Van Genuchten, M. Th. A comparison of numerical solutions of the one-dimensional unsaturated-saturated flow and mass transport equations. *Adv. Water Resour.*, 1982, **5** 47-55.

46. Van der Vorst, H. A. Bi-CGSTAB: A fast and smoothly convergent variant of BiCG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 1992, **13** 631-644.

47. Vogel, T., Huang, K., Zhang, R. & van Genuchten, M. Th. 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, Calif., 1996.

48. Webb, S. W. Generalization of Ross' tilted capillary barrier diversion formula for different two-phase characteristic curves. *Water Resour. Res.*, 1997, **33**(8) 1855-1859.

Appendix A

Jacobian J^Ψ for the pressure head Ψ as primary variable

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

$$J_{IJ}^{\psi}(\Psi_\tau^{n+1}, s_\tau^{n+1}) = \frac{\partial R_I^{n+1}(\Psi_\tau^{n+1}, s_\tau^{n+1})}{\partial \psi_{\tau J}^{n+1}}$$

$$= J_{IJ}^{\psi 1} + J_{IJ}^{\psi 2} + J_{IJ}^{\psi 3} + J_{IJ}^{\psi 4} - J_{IJ}^{\psi 5}$$

$$= \frac{\partial O_{IL}(\Psi_\tau^{n+1})}{\partial \psi_{\tau J}^{n+1}}\left[\frac{\sigma}{\Delta t_n}(\psi_{\tau L}^{n+1} - \psi_L^n) - (\sigma - 1)\dot{\psi}_L^n\right] +$$

$$+ B_{IL}\frac{\partial s_{\tau L}^{n+1}}{\partial \psi_{\tau J}^{n+1}}\frac{\sigma}{\Delta t_n} \qquad\qquad (A1)$$

$$+ \psi_{\tau L}^{n+1}\frac{\partial K_{IL}(s_\tau^{n+1})}{\partial \psi_{\tau J}^{n+1}}$$

$$+ \frac{\sigma O_{IJ}(s_\tau^{n+1})}{\Delta t_n} + K_{IJ}(\Psi_\tau^{n+1})$$

$$- \frac{\partial F_I(s_\tau^{n+1})}{\partial \psi_{\tau J}^{n+1}}$$

The partial Jacobians in eqn (A1) are obtained as follows

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

no summation over I and J (A2)

$$J_{IJ}^{\psi 2} = \sum_e \int_{\Omega_e} N_I \varepsilon C_{\tau J}^{n+1}\delta_{IJ}\frac{\sigma}{\Delta t_n} \qquad (A3)$$

no summation over I and J

$$J_{IJ}^{\psi 3} = \sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i}K_{ij}N_J G_{\tau J}^{n+1}\frac{\partial N_L}{\partial x_j}\psi_{\tau L}^{n+1} \qquad (A4)$$

no summation over J

$$J_{IJ}^{\psi 4} = \frac{\sigma O_{IJ}(s_\tau^{n+1})}{\Delta t_n} + K_{IJ}(\Psi_\tau^{n+1}) \qquad (A5)$$

$$J_{IJ}^{\psi 5} = -\sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i}K_{ij}N_J(1 + \chi)G_{\tau J}^{n+1}e_j \qquad (A6)$$

no summation over J

with

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

and

$$G_{\tau J}^{n+1} = \frac{\partial K_r(\psi_{\tau J}^{n+1})}{\partial \psi_{\tau J}^{n+1}} \qquad (A8)$$

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 (1-3)-(1-6) or numerically from chord slope approximations (Appendix C) for the known variables s and ψ at the iterate τ, the node J and the time plane $n + 1$. Here, C_τ^{n+1} is the moisture capacity function known from the standard unsaturated modeling.

Appendix B

Jacobian J^s for the saturation s as primary variable

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

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

The partial Jacobians in eqn (B1) are obtained as follows

$$
J_{IJ}^{s1} = \sum_e \int_{\Omega_e} N_I S_o \delta_{IJ} \left[\frac{\sigma}{\Delta t_n}(\psi_{\tau J}^{n+1} - \psi_J^n) - (\sigma - 1)\dot{\psi}_J^n \right] \tag{B2}
$$

no summation over I and J

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

$$
J_{IJ}^{s3} = \sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i} K_{ij} N_J G_{\tau J}^{n+1} \widehat{C}_{\tau J}^{n+1} \frac{\partial N_L}{\partial x_j} \psi_{\tau L}^{n+1} \tag{B4}
$$

no summation over J

$$
J_{IJ}^{s4} = \left(\frac{\sigma O_{IJ}(s_\tau^{n+1})}{\Delta t_n} + K_{IJ}(\Psi_\tau^{n+1}) \right) \widehat{C}_{\tau J}^{n+1} \tag{B5}
$$

no summation over J

$$
J_{IJ}^{s5} = -\sum_e \int_{\Omega_e} \frac{\partial N_I}{\partial x_i} K_{ij} N_J (1 + \chi) G_{\tau J}^{n+1} \widehat{C}_{\tau J}^{n+1} e_j \tag{B6}
$$

no summation over J

with the inverse moisture capacity

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

which can be either derived analytically from eqns (1-3) and (1-5) or numerically by using chord slope approximations (Appendix C). Notice, it is necessary to use the pressure head ψ instead of the hydraulic

head h to evaluate the moisture capacity functions $C_{\tau J}^{n+1}$ and $\widehat{C}_{\tau J}^{n+1}$. 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 $\widehat{C}_{\tau J}^{n+1}$ is not simply invertible for h because $\partial \psi/\partial s = \partial h/\partial s - \partial z/\partial s$.

Appendix C

Chord slope approximations of saturation derivatives

In contrast to analytical derivatives in form of the moisture capacity C_τ^{n+1} (A7) and its inverse $\widehat{C}_\tau^{n+1}$ (B7) chord 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 (1-35), (1-36) for the current time plane $n+1$. For instance, a simple first-order accurate finite difference approximation of C_τ^{n+1} would lead to

$$C_{\tau I}^{n+1} = \frac{s_{\tau I}^{n+1} - s_I^n}{\psi_{\tau I}^{n+1} - \psi_I^n} \quad \text{(C1)}$$

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

$$C_{pI}^{n+1} = \frac{s_{pI}^{n+1} - s_I^n}{\psi_{pI}^{n+1} - \psi_I^n} \quad \text{(C2)}$$

It can be easily seen that this derivative is nothing more than the quotient of the acceleration vectors (1-35) for the saturation and the pressure head

$$C_{pI}^{n+1} = \frac{\dot{s}_I^n}{\dot{\psi}_I^n} \quad \text{(C3)}$$

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

A corresponding second-order accurate chord slope approximation suited for the TR scheme can be similarly derived using eqn (1-41) as

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

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

Note here that limitations exist for the chord slope approximations if the denominator of eqns (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.

1. On the primary variable switching technique for simulating unsaturated-saturated flows

Treatment of free surfaces in 2D and 3D groundwater modeling

2

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

Nomenclature

ABSTRACT

The present paper aims at a discussion of the numerical requirements and efforts for the finite-element modeling of transient free surface(s) flow and transport problems in two and three dimensions. Following aspects are emphasized: (1) Classic groundwater against unsaturated-saturated modeling approaches. (2) Moving mesh approaches *versus* fixed grid strategies for multi-layered aquifers with their advantages and drawbacks. (3) Generalized boundary and constraint conditions for flow and transport modeling needed for a free-surface analysis. (4) Introducing the BASD (Best-Adaptation-to-Stratigraphic-Data) technique as a new numerical strategy to automatically adapt the location of water table(s) to all relevant data of a hydro-stratigraphic initial structure with parameter discontinuities. (5) Theoretical basis of a pseudo-unsaturated modeling approach and its limitation.

The impact of the numerical approaches is studied for selected applications: *(i)* benchmarking moisture dynamics in homogeneous and layered soils, *(ii)* drainage experiment, *(iii)* dam seepage modeling, *(iv)* benchmarking the mine flooding for a generic 3D pit geometry, and *(v)* real-site modeling of complex flow and contaminant transport problems.

Key words: groundwater, free-surface problems, unsaturated-saturated porous media, finite-element method, mesh adaptation, moisture movement, dam seepage, mine flooding

Latin symbols

A		curve-fitting parameter;
B		curve-fitting parameter;
C, C_o	ML^{-3}	concentration and reference concentration of a miscible chemical species, respectively;
$C(\psi)$	L^{-1}	moisture capacity storage;
c^f, c^s	$L^2 T^{-2} \Theta^{-1}$	specific heat capacity of fluid and solid, respectively;
$\boldsymbol{D}$	$L^2 T^{-1}$	tensor of mechanical dispersion;
D_d	$L^2 T^{-1}$	molecular diffusion in the porous medium;
$\boldsymbol{e}$	l	gravitational unit vector;
G		quantity;
g	LT^{-2}	gravitational acceleration;
h	L	hydraulic (piezometric) head;
$\boldsymbol{I}$	l	unit tensor;
$\boldsymbol{K}$	LT^{-1}	tensor of hydraulic conductivity for the saturated medium (anisotropy);
$\boldsymbol{k}$	L^2	tensor of permeability for the saturated medium (anisotropy);

2. Treatment of free surfaces in 2D and 3D groundwater modeling

K_r	1	relative hydraulic conductivity ($0 < K_r \leq 1$, $K_r = 1$ if saturated at $s^f = 1$);
m	1	$1 - 1/n$ curve-fitting parameter (Mualem assumption);
N		number of intersections;
N_i		finite element shape function at node i;
$\boldsymbol{n}$	1	normal unit vector (positive outward);
n	1	$n > 1$, pore size distribution index, approximately in the range $1.25 < n < 6$;
p^f	$ML^{-1}T^{-2}$	fluid pressure;
Q_h	T^{-1}	fluid flow sink/source;
Q_C	$ML^{-3}T^{-1}$	bulk mass sink/source;
Q_T	$ML^{-1}T^{-3}$	bulk thermal sink/source;
Q_T^f, Q_T^s	$ML^{-1}T^{-3}$	fluid and solid thermal sink/source, respectively;
$\boldsymbol{q}$	LT^{-1}	Darcy flux vector;
q_n		normal flux on a boundary (positive outward);
R, R_d	1	retardation and derivative retardation, respectively;
R_o	LT^{-1}	infiltration or evaporation rate on a free surface;
S		areal property;
S_o	L^{-1}	storage coefficient;
s^f	1	saturation of the fluid phase ($0 < s^f \leq 1$, $s^f = 1$ if medium is saturated);
s_e^f	1	effective saturation of fluid;
s_r^f	1	residual saturation of fluid;
s_s^f	1	maximum saturation of fluid;
T, T_o	Θ	temperature and reference temperature, respectively;

$\boldsymbol{x}$	L	Eulerian spatial coordinate vector;
z	L	elevation above a reference datum;
z_s^n	L	elevation vector for slice s and at time level n;

Greek symbols

α	L^3M^{-1}	solutal expansion coefficient;
α		curve-fitting parameter;
β	Θ^{-1}	thermal expansion coefficient;
β		curve-fitting parameter;
β_L, β_T	L	longitudinal and transverse dispersivity, respectively;
Γ		boundary;
γ	L^{-1}	fluid compressibility;
Δt_n	T	time increment at level n;
δ_{ij}		Kronecker operator;
ε	1	porosity ($0 < \varepsilon \leq 1$);
ε_e	1	effective porosity at the free surface;
θ^f	1	$s^f\varepsilon$, volumetric moisture content ($0 < \theta^f \leq \varepsilon$, $\theta^f = \varepsilon$ if medium is saturated);
θ_r^f	1	residual volumetric moisture content;
θ_s^f	1	saturated volumetric moisture content;
ϑ	T^{-1}	chemical decay rate;
κ	1	curve-fitting parameter;
$\underline{\Delta}^f$	$MLT^{-3}\Theta^{-1}$	tensor of thermal hydrodynamic dispersion of fluid phase;
λ^f, λ^s	$MLT^{-3}\Theta^{-1}$	thermal conductivity for fluid and solid, respectively;
μ^f, μ_o^f	$ML^{-1}T^{-2}$	dynamic viscosity and reference dynamic viscosity of fluid, respectively;

ρ^f, ρ_o^f	ML^{-3}	fluid density and reference fluid density, respectively;
ρ^s	ML^{-3}	solid density;
Υ	T	final time;
ϕ	L^{-1}	skeleton compressibility;
χ		adsorption function to describe Henry, Freundlich and Langmuir isotherms;
ψ	L	pressure head ($\psi > 0$ saturated medium, $\psi \leq 0$ unsaturated medium);
Ω		domain;

Subscripts

a	air entry;
i	property index
i, j, k	nodal indices;
l	intersected layer;
l	direction of gravity in the Cartesian coordinate system;
o	reference value;
s	slice index;
t	time-dependence;

Superscripts

d	number of space dimensions;
e	element;
f	fluid (water) phase;
n	time level;
s	solid phase;

2.1 Introduction

Subsurface flow and transport phenomena involving free surface(s) represent a general and important class of nonlinear problems. In the past, various approaches and computational methods have been established for solving groundwater free-surface problems in two and three dimensions with more or less success. Most of them are constrained with respect to the range of application and the practical motivation, for instance, local-scale dam seepage problems, regional groundwater flow modeling or moisture movement in the vadose zone for soil columns have led to quite different solution strategies[1-19]. Generally, in modeling free-surface problems two conceptual models can be chosen: (1) the *unsaturated-saturated modeling approach* and (2) the *fully saturated, water-table, classic groundwater modeling approach* (*cf.* Fig. 2.1). Each of them has their advantages and drawbacks. While the unsaturated-saturated approach involves the inclusion of the entire flow domain in the analysis, the fully saturated approach considers only the domain below the free surface where the water table is treated as a moving material interface.

From the physical point of view the unsaturated-saturated modeling approach provides the most rigorous treatment of computing free-surface problems. However, its solution enlarges the computational effort and has to tackle the strong nonlinearities in the governing equations for flow and transport. In the numerical solution process convergence problems can occur, especially under dry conditions. Furthermore, in many engineering applications the data of unsaturated characteristics are often not available. For a specific site the initial saturation states, capillary pressure and relative

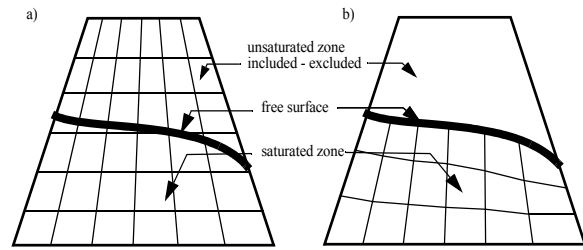conductivity relationships are data which are often difficult to obtain.

Figure 2.1 a) Unsaturated-saturated modeling approach using a fixed (invariant) mesh, b) fully saturated classic water-table modeling approach with a moving (variable) mesh.

On the other hand, common groundwater water-table models suffer from a number of limitations. The conception of free surfaces in groundwater can fail for complex applications, e.g., if there is no coherent water table and free surfaces become highly dynamic and isolated (trapped or perched water) which often happens if the water table has a deep location or water utilization (and drainage) in an aquifer system lead to partly dewatered regions while the layers above remains saturated. Such processes combined with tasks of saltwater intrusion or geothermal supply are very difficult to solve without an unified unsaturated-saturated flow approach. Otherwise, mesh adaptation due to free-surface movement is difficult to control numerically, especially for complex aquifer systems with high parameter contrasts and sharp zones of water depression. On the other hand, the use of fixed (unmoved) meshes in a conventional groundwater analysis can arise problems if (a) parts of the aquifer fall 'dry' because the handling

of such dry mesh cells is often heuristic and physically 'incorrect' and (b) mass transport is to be modeled because contaminants can be 'frozen' in 'dry' cells instead of moving according to the water table.

The majority of today's groundwater models for free surfaces (unconfined, phreatic aquifers) deals with the fully saturated zones only. Accordingly, it should seem to be a standard and well-solved task. However, the practical modeler can still report on a number of lacks and troubles under real applications. The major difficulties refer to the following: (1) The free-surface problem is mostly solved only in a non-rigorous manner, i.e., the kinematic boundary conditions are adapted by *ad-hoc* approaches (e.g. by introducing an auxiliary 'well-term') such as done in the widely used package MODFLOW[20]. Criticisms were recently summarized by Yeh *et al.*[16] and Knupp[17]. While Yeh[16] modeled homogeneous 3D domains for which a moving technique is much simpler, Knupp[17] developed an improved moving grid technique for a finite volume code which allows the computation of regional situations at complex stratigraphy and heterogeneous conditions. However, its proposed algorithm permits motion of only the upper portion of the grid. (2) Special handling is needed if parts of the domain intermediately become dry. There are different 'tricks' to overcome such situations (e.g., frozen cells, converting procedures, intermediate deletion of elements). Here is a great influence of 'dubious' manipulations in 'free' computer codes. Accordingly, more general techniques are required to attain robust, balance-accurate and non-oscillatory solutions. (3) Multiple (more than one) free surfaces in an aquifer system are often difficult to tackle. The storage coefficients in the layered system become strongly dependent on the dynamically wetted

element conditions. (4) The existence of free-surface conditions in the context of contaminant mass or heat transport processes, including density effects, forces to a generalization of the solution strategy.

The present paper aims at an unified handling of free-surface problems both in an unsaturated-saturated approach and in the fully saturated modeling approach embodied in one code to get more flexibility and robustness in the numerical solution. The basic formulation is based on the balance equations for flow, contaminant mass and heat transport in unsaturated-saturated media with density and viscosity coupling effects. Related boundary conditions and boundary constraints are discussed for free surfaces, seepage faces, ponding and drainage boundaries. The free surface and seepage conditions represent kinematic formulations which lead to a rigorous approach to solve the flow and transport equations in the fully saturated groundwater domain. For these purposes a more general moving mesh strategy, called BASD, will be introduced which is capable of computing movable finite element meshes in three dimension even under general stratigraphic heterogeneous material conditions. The moving mesh technique is compared with the conventional fixed grid technique for saturated flow and the more general unsaturated-saturated modeling approach.

The motivation for the developments is in complex 3D and 2D flow and transport problems in mining and water management problems. The impact of the used strategies will be shown along the moisture movement in unsaturated homogeneous and layered soils, drainage and seepage problems, a generic problem of pit flooding and in modeling of real-site mining problems.

All developments and computations refer to the interactive groundwater simulation system FEFLOW[21].

2.2 Governing Equations

The governing equations for the unsaturated-saturated flow, contaminant mass and heat transport are derived from the macroscopic phase-related conservation principles for mass, linear momentum and energy[22]. Let $\Omega_t \subset \mathfrak{R}^d$ and $(0, \Upsilon)$ be the spatial and temporal domains respectively, where d is the number of space dimensions (2 or 3), and let Γ_t denote the boundary of Ω_t. The subscript t implies the time dependence of the spatial domain (if the domain becomes time-invariant: $\Omega_t \to \Omega$ and $\Gamma_t \to \Gamma$). The spatial and temporal coordinates are denoted by $\boldsymbol{x} \in \Omega_t$ and $t \in (0, \Upsilon)$. The following nonlinear system finally results which has to be solved in two and three dimensions:

$$[S_o \cdot s^f(\psi) + \varepsilon \cdot C(\psi)]\frac{\partial h}{\partial t} + \nabla \cdot \boldsymbol{q} = Q_h \qquad (2\text{-}1)$$

$$\boldsymbol{q} = -K_r(s^f)\boldsymbol{K}f_\mu\left(\nabla h + \frac{\rho^f - \rho_o^f}{\rho_o^f}\boldsymbol{e}\right) \qquad (2\text{-}2)$$

$$s^f(\psi)R_d(C)\frac{\partial C}{\partial t} + \boldsymbol{q} \cdot \nabla C - \nabla \cdot [(\varepsilon s^f(\psi)D_d\boldsymbol{I} + \boldsymbol{D}) \cdot \nabla C] +$$
$$+ [s^f(\psi)R(C)\vartheta + Q_h]C = s^f(\psi)Q_C \qquad (2\text{-}3)$$

2. Treatment of free surfaces in 2D and 3D groundwater modeling

$$[s^f(\psi)\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s]\frac{\partial T}{\partial t} + \rho^f c^f \boldsymbol{q} \cdot \nabla T$$

$$- \nabla \cdot [(\underline{\Lambda}^f + (1-\varepsilon)\lambda^s \boldsymbol{I}) \cdot \nabla T] +$$

$$+ \rho^f c^f Q_h(T-T_o) = Q_T(\psi) \qquad (2\text{-}4)$$

with the definitions and constitutive relationships

$$h = \frac{p^f}{\rho_o g} + z = \psi + z \qquad (2\text{-}5a)$$

$$S_o = \varepsilon\gamma + (1-\varepsilon)\phi \qquad (2\text{-}5b)$$

$$C(\psi) = \frac{\partial s^f(\psi)}{\partial \psi} \qquad (2\text{-}5c)$$

$$\boldsymbol{K} = \frac{k\rho_o^f g}{\mu_0^f} \qquad (2\text{-}5d)$$

$$f_\mu = \frac{\mu_0^f}{\mu^f(C,T)} \qquad (2\text{-}5e)$$

$$\rho^f = \rho_o^f[1 + \alpha(C-C_o) - \beta(T-T_o)] \qquad (2\text{-}5f)$$

$$R(C) = \varepsilon + (1-\varepsilon)\chi(C) \qquad (2\text{-}5g)$$

$$R_d(C) = \varepsilon + (1-\varepsilon)\frac{\partial[\chi(C)\cdot C]}{\partial C}$$

$$\boldsymbol{D} = (\beta_L - \beta_T)\frac{\boldsymbol{q}\otimes\boldsymbol{q}}{\|\boldsymbol{q}\|} + \beta_T\|\boldsymbol{q}\|\boldsymbol{I} \qquad (2\text{-}5h)$$

$$\underline{\Lambda}^f = \rho^f c^f \boldsymbol{D} + \varepsilon s^f(\psi)\lambda^f \boldsymbol{I} \qquad (2\text{-}5i)$$

$$Q_T(\psi) = s^f(\psi)\varepsilon Q_T^f + (1-\varepsilon)Q_T^s \qquad (2\text{-}5j)$$

To solve the nonlinear equations (2-1) to (2-4) for h, $\boldsymbol{q}$, C, and T under unsaturated-saturated conditions constitutive relationships for the saturation-dependent moisture capacity $C(\psi)$ and hydraulic conductivity $K_r(s^f)$ have to be specified in form of empirical relationships for the capillary pressure head-saturation $\psi(s^f)$, with its inverse $s^f(\psi)$, and for the relative conductivities $K_r(s^f)$. The following parametric models will be preferred[23]:

(1) van Genuchten-Mualem model:

$$s_e^f = \begin{cases} \dfrac{1}{[1 + |A\psi|^n]^m} & \text{for} \quad \psi < 0 \\ 1 & \text{for} \quad \psi \geq 0 \end{cases} \qquad (2\text{-}6)$$

$$K_r = (s_e^f)^{\frac{1}{2}}\left\{1 - \left[1 - (s_e^f)^{\frac{1}{m}}\right]^m\right\}^2 \qquad (2\text{-}7)$$

with the effective saturation of fluid

$$s_e^f = \frac{s^f - s_r^f}{s_s^f - s_r^f} = \frac{\theta^f - \theta_r^f}{\theta_s^f - \theta_r^f} \qquad (2\text{-}8)$$

(2) Brooks-Corey model:

$$s_e^f = \begin{cases} \dfrac{1}{|\alpha\psi|^n} & \text{for} \quad \psi < -1/\alpha \\ 1 & \text{for} \quad \psi \geq -1/\alpha \end{cases} \qquad (2\text{-}9)$$

$$K_r = (s_e^f)^\kappa \qquad (2\text{-}10)$$

(3) Haverkamp model:

$$s_e^f = \begin{cases} \dfrac{\alpha}{\alpha + |\psi|^\beta} & \text{for} \quad \psi < 0 \\ 1 & \text{for} \quad \psi \ge 0 \end{cases} \qquad (2\text{-}11)$$

$$K_r = \begin{cases} \dfrac{A}{A + |\psi|^B} & \text{for} \quad \psi < 0 \\ 1 & \text{for} \quad \psi \ge 0 \end{cases} \qquad (2\text{-}12)$$

The above strongly nonlinear parametric curves describes the fluid saturation $s^f(\psi)$ and the relative hydraulic conductivity $K_r(s^f(\psi))$ as a function of the pressure head ψ in an unique manner. They are continuous over the entire range of ψ which is an useful feature for the numerical implementation. Under saturated conditions $s^f \equiv 1$ the nonlinearities in the parametric relationships vanish, however, for free-surface conditions a nonlinear boundary-value problem remains to be solved.

2.3 Initial, Boundary and Constraint Conditions

2.3.1 Initial conditions

The initial condition on the hydraulic head h, concentration C and temperature T are specified on Ω_0:

$$\left. \begin{aligned} h(\boldsymbol{x}, 0) &= h_0 \\ C(\boldsymbol{x}, 0) &= C_0 \\ T(\boldsymbol{x}, 0) &= T_0 \end{aligned} \right\} \qquad (2\text{-}13)$$

It is obvious due to the above parametric relationships the initial hydraulic head distributions represents at the same time an alternate initial distribution for the pressure head ψ, fluid saturation s^f and the moisture content θ^f:

$$\left. \begin{aligned} \psi(\boldsymbol{x}, 0) &= \psi_0 \\ s^f(\boldsymbol{x}, 0) &= s_0^f \\ \theta^f(\boldsymbol{x}, 0) &= \theta_0^f \end{aligned} \right\} \qquad (2\text{-}14)$$

2.3.2 Standard boundary conditions

Denoting Γ_t^1 and Γ_t^2 two disjoint portions of the total boundary $\Gamma_t = \Gamma_t^1 \otimes \Gamma_t^2$, the following formulations for boundary conditions (BC's) are specified for the Eqn. (2-1) to (2-4),

$$\left. \begin{aligned} G &= G_1 & \text{on} & \quad \Gamma_t^1 \\ q_n^G + a(G_2 - G) &= b & \text{on} & \quad \Gamma_t^2 \end{aligned} \right\} \qquad (2\text{-}15)$$

with the normal fluxes

$$q_n^G = \begin{cases} -\left[K_r(s^f)Kf_\mu \cdot \left(\nabla h + \dfrac{\rho^f - \rho_o^f}{\rho_o^f}e\right)\right] \cdot n \\ \\ -[(\varepsilon s^f(\psi)D_d I + D) \cdot \nabla C] \cdot n \\ \\ -[(\Delta^f + (1-\varepsilon)\lambda^s I) \cdot \nabla T] \cdot n \end{cases}$$ (2-16a)

$$G = \begin{cases} h \\ C \\ T \end{cases}$$ (2-16b)

where on Γ_t^1 Dirichlet BC occurs and on Γ_t^2 a more general form of a Robin type BC is imposed. If $a = 0$ a Neumann flux-type BC of 2nd kind results, while for $b = 0$ a common Cauchy BC of 3rd kind is given. In (2-16a) n corresponds to the normal unit vector (positive outward), G_1 and G_2 are prescribed boundary values for h, C, or T on Γ_t^1 and Γ_t^2, respectively. In the present context some specifications of the boundary conditions are becoming important which are described next.

2.3.3 'Drainage' gradient-type boundary conditions

Applied to unsaturated problems a Neumann ($a = 0$) flux-type boundary condition in the form

$$-\left[K_r(s^f)Kf_\mu \cdot \left(\nabla h + \dfrac{\rho^f - \rho_o^f}{\rho_o^f}e\right)\right] \cdot n = q_n^h$$ (2-17)

can be sometimes inappropriate, for instance if model-

ing a drainage boundary in the vadose zone with a bottom outflow boundary condition for situations where the water table is situated far below the domain of interest. Here, a gradient-type boundary condition is often to be preferred[24]:

$$-\left[Kf_\mu \cdot \left(\nabla h + \dfrac{\rho^f - \rho_o^f}{\rho_o^f}e\right)\right] \cdot n = q_n^{hgrad}$$ (2-18)

On such a boundary it can be assumed the pressure gradient diminishes $\nabla\psi \approx 0$ and since $\nabla h = \nabla\psi + \nabla z$, Eq. (2-18) is practically applied in the following form:

$$-\left[Kf_\mu \cdot \left(\nabla z + \dfrac{\rho^f - \rho_o^f}{\rho_o^f}e\right)\right] \cdot n = q_n^{hgrad}$$ (2-19)

Once $\nabla z \cdot n \neq 0$ the boundary freely drains the flow domain due to the influence of gravity.

2.3.4 Free surface boundary conditions

The incorporation of free surface conditions is performed via proper kinematic boundary conditions. Starting point represents the conservation relationship at a macroscopic surface of discontinuity[22], the interfacial free surface. It leads finally to the formulation:

$$\left.\begin{aligned} \left(R_o - \varepsilon_e \dfrac{\partial h}{\partial t}\right)n_l &= -q_n^h \\ \\ h &= z \end{aligned}\right\}$$ (2-20)

where n_l corresponds to the component of the unit outward normal vector $\boldsymbol{n}$ directed along the gravity direction (z-coordinate). As seen from Eq. (2-20) along a free surface two boundary conditions are to be satisfied simultaneously:

- a prescribed flux rate (as an infiltration or, if equal to zero, then impervious) and
- the location corresponds to the hydraulic head (constant pressure level)

which leads to a nonlinear boundary-value problem because the location of a free surface is initially unknown.

2.3.5 Seepage face boundary conditions

A seepage face represents a specific free surface condition for which the geometry as a part of the boundary of the flow domain Ω_t is known, except for the location of its end points[23]. It reduces to a Dirichlet-type condition of a prescribed hydraulic head condition in the form

$$h = z \tag{2-21}$$

Since the range of the seepage face is initially unknown its solution also leads to a nonlinear task. Practically, it can be solved by applying additional constraint conditions on the boundary which will be described further below. Commonly, a seepage face condition is imposed on such portions where the balanced flux going through the boundary is directed outward (it means the seepage face exists only there where the boundary drains the flow domain).

The seepage face condition $h = z$ can also be considered as a reference head condition outside the flow domain. As an alternate to (2-21) it can act as a Cauchy-type condition in the form

$$q_n^h = -a(z - h) \tag{2-22}$$

where the transfer coefficient a mimics a 'resistance' to control the outflow through the seepage face (e.g., at a dam covering).

2.3.6 Surface ponding boundary condition

Surface ponding denotes a 'surface reservoir' boundary condition of the type

$$\left(R_o - \frac{\partial h}{\partial t}\right)n_l = -q_n^h \tag{2-23}$$

which is similar to free-surface condition as situated in an 'air layer' for what the effective porosity becomes unity ($\varepsilon_e = 1$). This condition permits water to build up on the surface. The height of the surface water layer increases or reduces according to the rate R_o.

Alternatively, if the surface reservoir is deemed infinity the boundary of the surface can be imposed by a prescribed water table condition $h = z$ as long as the infiltration is associated with a water table above a given elevation. This can be easily solved in combination with boundary constraints where a Neumann-type

BC is constrained by min-max hydraulic head bounds in form of Dirichlet-type conditions.

2.3.7 Constraints of boundary conditions

Constraints are limitations for all types of boundary conditions. They result from the requirement that boundary conditions should only be valid as long as minimum and maximum bounds are satisfied. If during a simulation run the conditions are exceeded or fall below, the constraints are to be assigned as new intermediate boundary conditions. As already indicated above, seepage face conditions belong to this type: A Dirichlet-type BC $h = z$ is imposed along a boundary on which a seepage face can potentially occur. Additionally, this condition is committed to the constraint that the flux has to release (drain) the flow domain. In such a procedure the flux through the boundary has to be continuously checked (e.g., is Q becoming positive or negative) to decide on the acceptance of the original boundary condition or on its intermediate replacement by appropriate flux-type boundary conditions.

The formulation of constraints is based on the formalism of *complementary conditions* for a type of boundary condition. Accordingly, a potential condition (hydraulic head, concentration or temperature) is constrained by maximum and minimum flux relations (fluid, mass or heat fluxes). On the other hand, flux relationships are constrained by complementary potential limits, i.e., the fluid flux is constrained by min-max heads, the contaminant mass flux by min-max concentrations, and the heat flux by min-max temperatures (for more details see Diersch[22]).

For instance, the minimum and maximum constraints of a Dirichlet-type concentration will lead to additional conditions in the following form (it reads: the imposed boundary condition $C = C_1(t)$ is accepted only if the related mass balance flux Q_C (and the related hydraulic head h_1) is within given min-max bounds, if not, these bounds have to be used as new boundary conditions, where the boundary type has to be changed from a Dirichlet-type into a flux-type boundary condition of a point sink/source Q_C)

$$
\text{1st kind} \quad C_1(t) \;\; \text{only if} \left\{ \begin{array}{c} Q_C^1 < Q_C^{max_1}(t) \\ \text{and} \\ Q_C^1 > Q_C^{min_1}(t) \\ \text{and} \\ h^{min_1} \le h_1 \le h^{max_1} \end{array} \right\}
$$

else set Q_C as an intermediate flux-type condition according to: $\qquad$ (2-24)

$$
Q_C = \left\{ \begin{array}{ll} Q_C^{max_1}(t) & \text{if} \quad \left\{ Q_C^1 \ge Q_C^{max_1}(t) \quad \text{and} \quad h^{min_1} \le h_1 \le h^{max_1} \right\} \\ Q_C^{min_1}(t) & \text{if} \quad \left\{ Q_C^1 \le Q_C^{min_1}(t) \quad \text{and} \quad h^{min_1} \le h_1 \le h^{max_1} \right\} \\ 0 & \text{if} \quad \{ h_1 < h^{min_1} \quad \text{or} \quad h_1 > h^{max_1} \} \end{array} \right.
$$

where Q_C^1 is the mass balance flux at the boundary point to be computed while the C_1 condition is imposed, $Q_C^{max_1}$ and $Q_C^{min_1}$ denote the prescribed time-dependent maximum and minimum bounds, respectively, and Q_C represents a singular mass sink/source to be set at the boundary point (node) instead of the original Dirichlet-type boundary condition. Similar expressions exist for the other types of boundary conditions. This procedure allows the control of concentra-

tion at the boundary in dependence on both the balanced flow conditions through the boundary (e.g., $Q_C^{min_1} \equiv 0$) and the location of possible free-surface conditions within the bounds h^{min}, h^{max}. The latter is very important for complex mine flooding processes as studied by Diersch et al.[25].

2.4 Finite Element Formulations

The finite element method (FEM) is used to solve the governing balance equations (2-1) to (2-4) with their constitutive relationships (2-5a) to (2-12) and the accompanying initial conditions (2-14), (2-15) and boundary conditions (2-15) to (2-24) for both the unsaturated-saturated approach and the fully saturated approach. In the general case the spatial finite-element discretization yields the following highly nonlinear matrix system:

$$O(h)\dot{h} + S(h)h = F(h, q, C, \dot{C}, T, \dot{T})$$
$$A(h)q = B(h, C, T)$$
$$P(h, C)\dot{C} + D(q, h, C, T)C = R(h, C)$$
$$U(h)\dot{T} + L(q, h, C, T)T = W(h, T)$$

$$(2\text{-}25)$$

where h, q, C, and T represent the resulting vectors of nodal hydraulic head, Darcy fluxes, contaminant concentration and temperature, respectively. Notice, the hydraulic head h is strictly correlated with the pressure head ψ and the fluid saturation s^f via the above definition and the constitutive equations. The superposed dot means differentiation with respect to time t. The matri-

ces S, A, O, P, and U are symmetric and sparse, while D and L are unsymmetric, however sparse too. The remaining vectors F, B, R and W encompass the right-hand sides of equations. The main nonlinear functional dependence is shown in parenthesis.

In solving the equation system for unsaturated-saturated conditions the domain and the boundary can be considered invariant, i. e., the used mesh becomes fixed and independent of time: $\Omega_t \rightarrow \Omega$, $\Gamma_t \rightarrow \Gamma$. This represents a great advantage, but it is paid by an enlarged solution domain and the strongly nonlinear dependences which increases the computational effort, partly dramatically.

Figure 2.2 indicates how the solution process is cycled for each equation inside the time loop if constraint conditions occur. The proposed adaptive strategy allows an accurate and non-oscillatory solution even for ill-posed problem formulations.

For solving the resulting large sparse matrix systems appropriate iterative solvers for symmetric and unsymmetric equations are applied. For the symmetric positive definite flow equations the conjugate gradient (CG) method is successful provided a useful preconditioning is applied. Standard preconditioner such as the incomplete factorization (IF) technique and alternatively a modified incomplete factorization (MIF) technique based on the Gustafsson algorithm are used. Different alternatives are available for the CG-like solution of the unsymmetric transport equations: a restarted ORTHOMIN (orthogonalization-minimization) method, a restarted GMRES (generalized minimal residual) technique and Lanczos-type methods, such as CGS (conjugate gradient square), BiCGSTAB (bi-conjugate gradient stable) and BiCGSTABP (postconditioned bi-conjugate gradient stable). For preconditioning an incomplete Crout decomposition scheme is applied. Commonly, BiCGSTABP is the first choice in the practical simulation of large problems.

Under saturated conditions the most of the h-dependencies in the matrix expressions vanish and the matrix system gets a much simpler form:
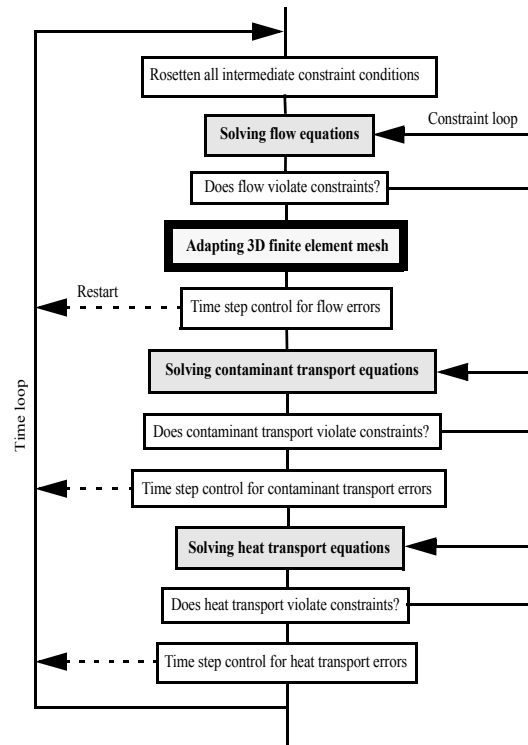
$$(2\text{-}26)$$

Figure 2.2 Adaptive strategy for transient free-surface flow, mass and heat transport.

The nonlinear equation system is temporarily discretized by predictor-corrector two-step marching schemes[26] of a first and second order accuracy in time. For the present nonlinear problems the Newton method is preferred for the unsymmetric transport equations, while Newton method can be omitted in the flow equations to maintain the symmetric property of the matrices.

$$O\dot{h} + Sh = F(h, q, C, \dot{C}, T, \dot{T})$$
$$Aq = B(C, T)$$
$$P(C)\dot{C} + D(q, C, T)C = R(C) \tag{2-27}$$
$$U\dot{T} + L(q, C, T)T = W(T)$$

where the remaining nonlinearities result from possible density and viscosity coupling effects. However, the free surface condition introduces a h-dependence in the solution domain $\Omega_t = \Omega_t(h)$ and the boundary $\Gamma_t = \Gamma_t(h)$. Accordingly, the mesh has to be adapted to the changed geometric relations during the simulation. This requires a moving mesh strategy which will be discussed further below.

In the preferred finite element method the free-surface boundary condition (2-20) is directly incorporated into the flow equations of (2-26) as written for the discretized 2nd order trapezoid rule

$$\left(\frac{2O}{\Delta t_n} + S\right)h^{n+1} = O\left(\frac{2}{\Delta t_n}h^n + \dot{h}^n\right) + F^{n+1}$$

$$O_{ij} = \sum_e \left(\int \int_{\Omega^e} \int S_o N_i N_j d\Omega + \int \int_{\Gamma_{fs}^e} \varepsilon_e N_i N_j n_l d\Gamma \right) \tag{2-28}$$

where n represents the time level. For saturated free-surface conditions the storage matrix O consists of two parts: a volume integral for the material compressibility and a surface integral for the storage of the material interfacial (fillable/drainable) property at the free surface.

In contrast to (2-28) the computation of the storage under unsaturated-saturated conditions is quite differ-

ent. Here the O-matrix holds the form:

$$O_{ij} = \sum_e \left(\int \int_{\Omega^e} \int S_o(s_k^f N_k)N_i N_j d\Omega + \int \int_{\Omega^e} \int \varepsilon\left(\frac{\partial s_k^f}{\partial \psi}N_k\right)N_i N_j d\Omega \right) \tag{2-29}$$

It reveals two volume integrals, where the second integral on the right-hand side involves the moisture capacity storage term which is highly nonlinear and dependent on the chosen capillary pressure head relationship as defined by the expressions (2-6), (2-9) or (2-11). Notice, for unsaturated conditions it has been shown[27,28] the consistent matrix O is to be transformed into a lumped matrix

$$O_{ij} = \sum_e \left(S_o \delta_{ij} \int \int_{\Omega^e} \int (s_k^f N_k)N_i d\Omega + \varepsilon \delta_{ij} \int \int_{\Omega^e} \int \left(\frac{\partial s_k^f}{\partial \psi}N_k\right)N_i d\Omega \right) \tag{2-30}$$

where here a row-sum (diagonal) scaling is preferred. It guarantees mass-conservative and nonoscillatory solutions.

2.5 BASD (Best-Adaptation-to-Stratigraphic Data)

For moving meshes in a fully saturated modeling approach an accurate and powerful technique is required to align and join the spatially-varying parameter fields according to the changeable free surface locations. Taking into consideration the parameters (e.g., conductivity) often have high contrasts it should be

clear a formal interpolation of the parameters onto the new mesh coordinates would lead to very poor solutions. The idea of BASD is quite simple: adapting the moving mesh as best as possible to the stratigraphic initial mesh to minimize parameter interpolation. Slices of the mesh are aligned in such a manner that the adjusted mesh is exactly fitted to parameter discontinu-ities if ever exist. Remaining slices can be shifted and repositioned to get a well-spaced nodal distribution in the depth without unnecessary mesh refinement and coarsening if attainable. The principle of the mesh adaptation process can be seen in Fig. 2.3.
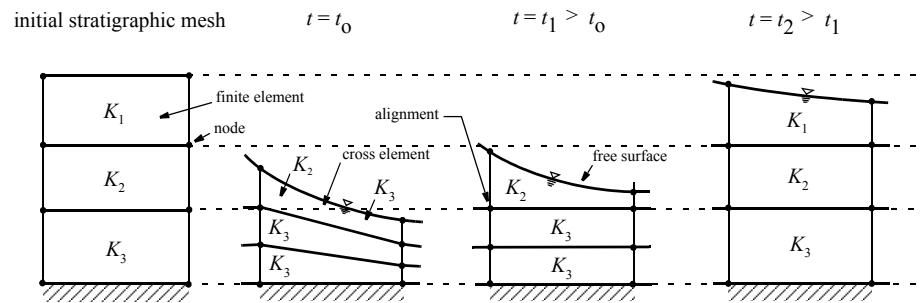

Figure 2.3 Moving mesh BASD technique of parameter adaptation applied to 3D free-surface problems: schematized example for a groundwater rise.

The initial stratigraphy consists of three layers with different conductivities. At the initial time t_o the water table is on a lower position. The mesh is accordingly shrunk where the lower two layers completely fit into the K_3 stratigraphy. However, the upper layer crosses between the K_2 and K_3 stratigraphy and a special treatment is here required. Such cross elements should be admitted only if unavoidable. A proper 3D interpolation technique has been developed which allows a data joining for elements intersecting an arbitrary number of stratigraphic layers as described below. If the water table ascends (Fig. 2.3 at time t_1) the moving mesh totally fits the K_2-K_3 stratum while the remaining slice is used to subdivide the widest nodal spacing,

here in the K_3 layer. At later time t_2 a further rise of the free surface occurs and the moving mesh slices appear to be well aligned to the data stratification without any need of interpolation.

The working steps of the BASD technique can be summarized as follows:

(A) Compute the hydraulic head $\boldsymbol{h}^{n+1}$ at the new time level according to Eq. (2-28).
(B) Determine a new free surface location for the upper slice $s = top = 1$ of the moving mesh

$$z_{top}^{n+1} = \boldsymbol{h}^{n+1}(x_i, t) \qquad (2\text{-}31)$$

satisfying the constant head condition of Eq. (2-20), where z_{top}^{n+1} corresponds to z-coordinates of the top slice.

(C) Adjust and distribute the inner slices, indexed by s, according to

$$z_s^{n+1} = z_{rig} + L^{rel}(z_s^n - z_{rig})$$
$$s = 2, ..., rig{-}1; \quad s = 1 \Rightarrow top \qquad (2\text{-}32)$$
$$L^{rel} = (z_{top}^{n+1} - z_{rig})/(z_{top}^n - z_{rig})$$

where z_{rig} is the firstly found *rigid* (nonmovable and time-independent) slice counting from top (at least the lowest slice describing the aquifer base is rigid) and L^{rel} the relation lengths. Special nesting rules have been developed as for the subdivision of overdue slices within layers enclosed by two rigid slices:

$$z_s^{n+1} = z_{s+1}^{n+1} + \frac{1}{n_d + n_h}(z_{rig}^{upper} - z_{rig}^{lower}) \qquad (2\text{-}33)$$

where z_{rig}^{upper} and z_{rig}^{lower} are the z-coordinates of the upper and lower rigid slice, respectively, n_d is the number of primary subdivisions and n_h the number of overdue (hanging) slices caused by slice shifting.

(D) Assign the parameter arrays according to the new layer positions. Two cases exist: (a) achievement of full alignment (no interpolation) and (b) data interpolation and joining for cross volume elements.

(E) Find out cross elements and join their properties. The joining process differs between volume-specified materials (such as conductivity, compressibility) and area-specified data (such as effective porosity). For volume-specified material data Gauss-Legendre

numerical integration is used to determine the partial volumes V_i of a finite element intersecting the stratigraphic contours. The material property K of the cross element is computed by a partial volume-weighted average as

$$K = \frac{1}{V}\sum_i V_i K_i \qquad (2\text{-}34)$$

instead of using a harmonic average, where i runs over the number of intersections N and K_i is the property of the intersected layer.

Similarly, a partial area-weighted averaging process is preferred for areal properties. However, it has been found a numerical integration is here insufficient. Analytical formulas have been developed to determine exactly the intersected areas of an element. It leads to a 'telescope' sum in the form

$$S = \sum_{i=1}^{l} (\lambda_{12}^i \lambda_{13}^i - \lambda_{12}^{i-1}\lambda_{13}^{i-1})S_i + (\lambda_{31}^l(1 - \lambda_{32}^l))S_l \qquad (2\text{-}35)$$
$$+ \sum_{i=l+1}^{N+1} (\lambda_{31}^{i-1}\lambda_{32}^{i-1} - \lambda_{31}^i \lambda_{32}^i)S_i$$

with the weights

$$\lambda_{mn}^i = \frac{1}{1 - \dfrac{h_n - z_n^i}{h_m - z_m^i}} \qquad i = 1, ..., N \qquad (2\text{-}36)$$

written for prismatic pentahedral elements, where S and S_i are the averaged and partial areal properties,

respectively, h_n and z_n correspond to the hydraulic head and the stratigraphic z-coordinates at node n, respectively, and the index l represents the lth intersected layer for which the partial area is not a triangle, generally a pentagon. Equivalent averages can be derived for hexahedral elements.

The use of the BASD technique for a complex multi-aquifer system can be seen in Fig. 2.4. It indicates how the mesh fits and moves through the complex stratigraphy consisting of a number of aquifers and aquitards.
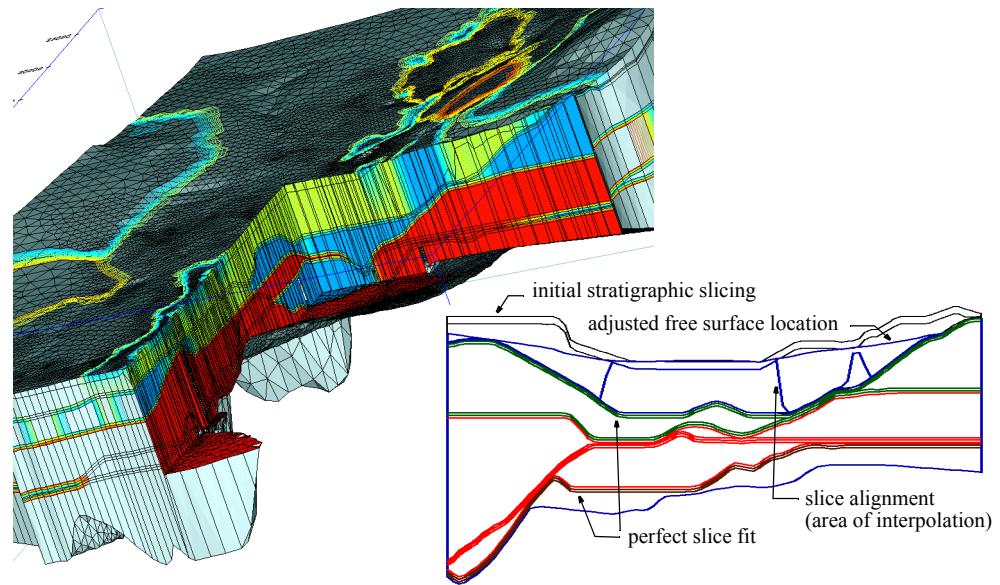

Figure 2.4 Moving mesh BASD strategy in a complex 3D stratigraphy (122,775 nodes and 226,394 elements): 3D model cut view (distribution of conductivities) and moving mesh along a cross section (initial stratigraphy *versus* adapted slicing).

2.6 Pseudo-Unsaturated Modeling Approach

Indeed, moving mesh strategies for adapting the free-surface location complicates the computational

process. Furthermore, if the free surface is not on the top position of the schematized aquifer system or if there are more than one free surface in the aquifer system (e. g., an additional free surface in a lower position) the problem cannot be solved alone on the basis of moving meshes. In these cases fixed mesh techniques become inevitable. It is a common practice in classic 3D groundwater modeling for flow in unconfined aquifers to use exclusively fixed grids (e.g.[20,29]). Fixed grid techniques have to mimic, more or less, unsaturated flow conditions to control the solution process for saturated, partly saturated or completely dry grid cells. Since a physically true unsaturated flow approach is avoided, such kind of unsaturated flow modeling represents only a physical approximation and quite different forms of implementation can be found in the literature (see discussion in Knupp[17]). Often, there is actually no rigorous physical basis in modifying the saturated flow conditions to achieve pseudo-unsaturated flows. Practically, the scaling of conductivity is used as a contrivance to obtain the solution in the saturated domain. For instance, the conductivity is assigned to a very small constant value as soon the pressure head becomes negative: $K/1000$ for $\psi < 0$ and K if $\psi \geq 0$ [30]. Apparently, this is a crude controlling procedure since it does not differ between the degrees of saturation of the elements. Desai and Li[11] have improved the technique for finite elements by introducing linear relationships of conductivity and storativity as function of the pressure head ψ. The linear functions operate as multipliers to the conductivity and storage terms ranging between maximum (saturated) and minimum (residual) factors.

The here proposed method is similar to Desai and Li[11], however, instead of prescribing an auxiliary linear pressure relationship the water (pseudo-)saturation computed for a finite element is used to 'down-scale' all balance terms in a natural way. The pseudo-saturation s_P^f is determined from the actual filling height of water in an element:

$$s_P^f(\psi) = \frac{d\Omega^f(\psi)}{d\Omega} \qquad (2\text{-}37)$$

Accordingly, the pseudo-saturation becomes related to the actual geometric condition of the used spatial discretization. It provides a geometry-consistent scaling of balance terms and has proven superior to a simple parameter-switching as stated above. An element e is considered saturated if $\psi \geq 0$ at all nodes of the element. Then it becomes $d\Omega^f = d\Omega$ and $s_P^f \equiv 1$. An element e is considered partially saturated (pseudo-unsaturated) if ψ changes its sign at the element nodes (e.g., $\psi < 0$ for the upper nodes and $\psi > 0$ for the lower (at least one) node(s). Then it is approximately $d\Omega^f \approx d\Gamma^f \cdot \psi_{lower}$. An element e is considered fully unsaturated (or dry) if $\psi < 0$ at all related element nodes. Since $d\Omega^f$ have to be positive the volume must be constrained by a minimum. Practically, a minimum filling height (e. g., 1 mm) is employed to limit $d\Omega^f$. This leads to a measure of a residual pseudo-saturation in such an element.

Using the expression (2-37) in the finite element equations (2-26) it leads to a natural approach for evaluating the corresponding integral terms in a weak solution. For instance, the part of the conductivity term of the flow equation takes the form:

$$\int\int\int_{\Omega^e}(\nabla N_i \cdot \boldsymbol{K} \cdot \nabla N_j)d\Omega^f \qquad (2\text{-}38)$$

$$= \int\int\int_{\Omega^e}(\nabla N_i \cdot \boldsymbol{K} \cdot \nabla N_j)s_P^f(\psi)d\Omega$$

$$= \int\int\int_{\Omega^e}(\nabla N_i \cdot K_r\boldsymbol{K} \cdot \nabla N_j)d\Omega$$

introducing a relative conductivity $K_r = s_P^f(\psi)$ as a linear function of ψ. Similarly, the storage term results as

$$O_{ij} = \sum_e\left(\int\int\int_{\Omega^e}S_o s_P^f(\psi)N_iN_jd\Omega + \int\int_{\Gamma^e_{fs\psi}}\varepsilon_e N_iN_jn_ld\Gamma\right) \qquad (2\text{-}39)$$

As seen the volume compressibility becomes a linear function of the pressure head ψ too. On the other hand, the surface integral which describes the kinematic free-surface boundary condition (2-20) is related to the geometric shape $\Gamma^e_{fs\psi}$ formed by the filling heights in the corresponding element (notice, while Γ^e_{fs} is the free surface facet which is built by an element (top) surface for moving meshes, accordingly Γ^e_{fs} is a part of the adapted boundary geometry itself; at fixed meshes $\Gamma^e_{fs\psi}$ lies in the interior of an element, accordingly the integral has to be evaluated for a surface which is spanned by the ψ heights).

It should be emphasized a pseudo-unsaturated modeling approach is suited to compute the location of a free surface, but, it is widely inappropriate to model a true unsaturated flow regime. The advantage is in its simplicity and robustness, but it is generally inferior to a moving mesh strategy with respect to the attainable accuracy.

2.7 Applications

2.7.1 Moisture dynamics in homogeneous and layered soils

First, let us study true unsaturated flows in both homogeneous and layered soils to show the capabilities and the requirements in simulating processes with variable saturation.

The first example refers to Celia *et al.*'s water infiltration problem[27] to benchmark the present solution of the unsaturated-saturated modeling approach for a strong infiltration front development. Celia *et al.* preferred a mixed $\theta - \psi$-form of the governing Richards equation to achieve accurate solution and sufficient mass conservation. Unlike, the present model is *h*-based and embedded in a second order predictor-corrector scheme for tackling the nonlinear solution process by an error-controlled timestep adaptation.

A detailed problem description of the Celia *et al.* problem is given elsewhere[27,31] and only the major characteristics are summarized: homogeneous soil column with a length of 1 m, van Genuchten-Mualem parametric model (2-6), (2-7) in using $n = 2$, $(m = 0.5)$, $A = 3.35$ 1/m, $\varepsilon = 0.368$, $s_r^f = 0.277$, and $s_s^f = 1.0$, saturated hydraulic conductivity of $0.922\cdot10^{-4}$ m/s, boundary condition of constant head $\psi = $ -0.75 m at the top and $\psi = $ -10.0 m at the bottom, and initial pressure head ψ_0 of -10.0 m. The same spatial discretization characteristics are applied as given in Celia *et al.*, who used $\Delta z = 0.5$ cm (dense grid) and $\Delta z = 2.5$ cm (coarse grid). Otherwise, Celia *et al.* used a constant time step length of $\Delta t = 60$ s for the dense-grid simulation.

A comparison with Celia *et al.*'s results gives very good agreements. Figure 2.5 shows the computed pressure profiles at time of 1 day for the case of the dense grid (notice, Celia *et al.*'s results are picked from a table presented in[31], where however only selected sample points are listed).
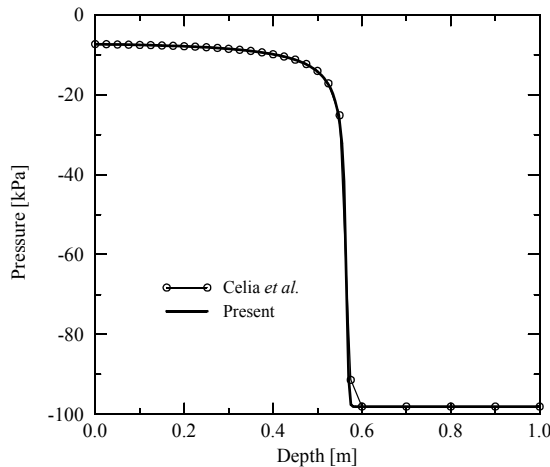

Figure 2.5 Pressure profiles at $t = 1$ day for the dense grid: Celia *et al.*'s $\theta - \psi$-based solution[27,31] using a mixed-form Picard iteration *versus* the present predictor-corrector h-based solution using an adaptive time stepping.

Celia *et al.* used both finite difference and finite element techniques. For the ψ-based Richards equation (which should be equivalent to a h-based form) they found nonoscillatory results only for finite differences, while finite elements echoed wiggles at the moisture front once coarse meshes were applied. They could improve the results if resorting to a modified Picard iteration technique based on a mixed $\theta - \psi$-form of the
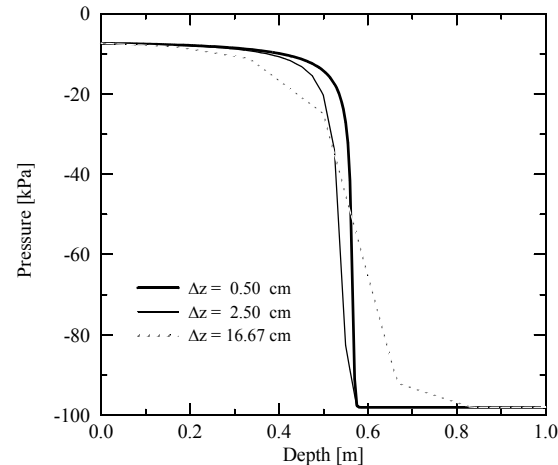
governing Richards equation.

Figure 2.6 Pressure profiles at $t = 1$ day for different grid spacing Δz computed by the present predictor-corrector technique with a h-based form of the Richards equation.

These difficulties as reported in[27] are completely prevented in the present predictor-corrector technique even for coarse meshes. Figure 2.6 displays the pressure profiles for different spatial discretizations. It reveals three things: Firstly, the proposed numerical technique gives wiggle-free solutions. The loss of accuracy against the dense grid solution is similar to that as observed by Celia *et al.* (for more details see[27]). Secondly, the accuracy of the solution is significantly influenced by the spatial (and temporal) discretization. That means, the simulation of unsaturated flow processes requires sufficiently refined meshes. Thirdly, a 50 % moisture profile centre point would not necessarily serve as an accurate identification of a possible free

surface location. As evidenced in Fig. 2.6, in dependence on the mesh refinement a leading or lagging phase error in the moisture profiles can occur, but, unfortunately, the effect of phase lag or lead does not appear as a simple function of grid spacing (compare also[27]).
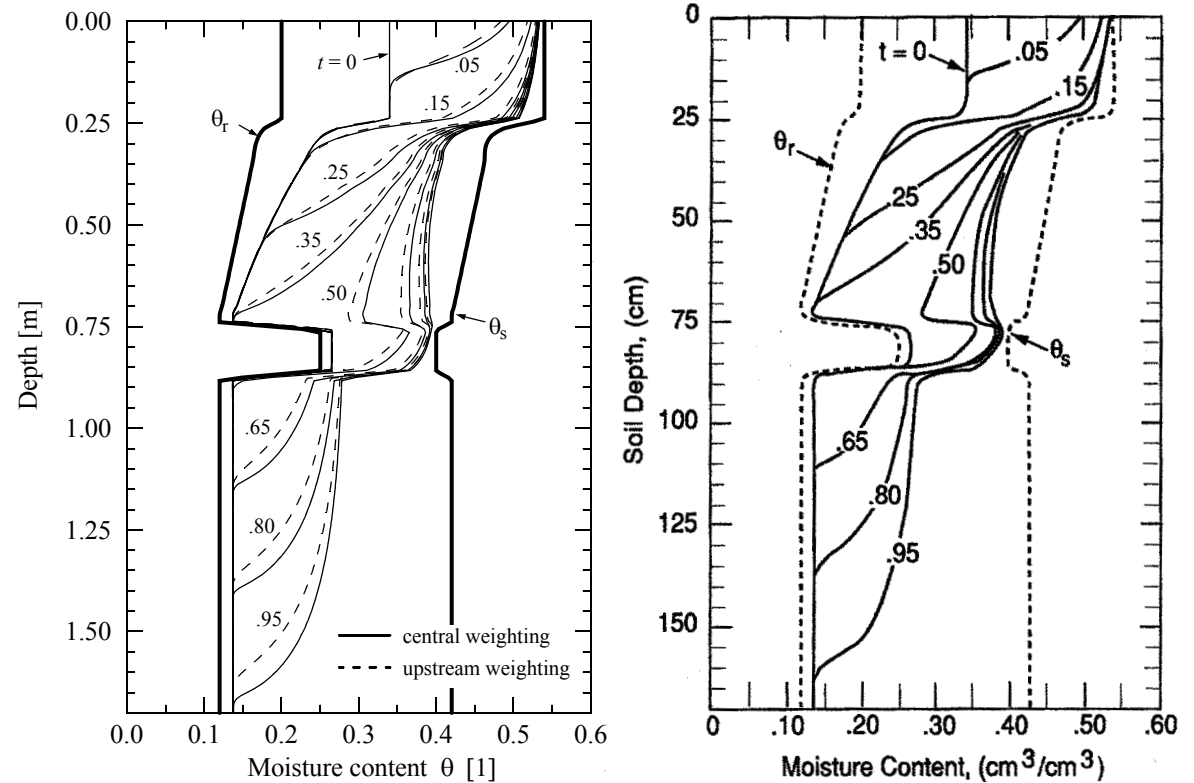

Figure 2.7 Simulated moisture-content profiles during infiltration: present solutions (left) and van Genuchten's results[15, 31] (right), time in days.

The second example concerns the simulation of the moisture movement in a layered soil as studied by van Genuchten[15]. 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 (87-170 cm), where the loamy-sand layer's properties change gradually with depth. The initial conditions for the flow are given by ψ_0 = -3.5 m. A constant flux is at the surface q_n^h = -0.25 m/d at $t \leq 1$ day (infiltration) and q_n^h = 0.005 m/d at $t > 1$ day (evaporation). At the bottom a drainage gradient-type BC (2-19) of $q_n^{hgrad} = -(-1) \cdot K|_{bottom}$ = 4 m/d is imposed, accordingly, the bottom boundary can freely drain. The parameter in the constitutive relations (van Genuchten-Mualem model) can be found in[31].
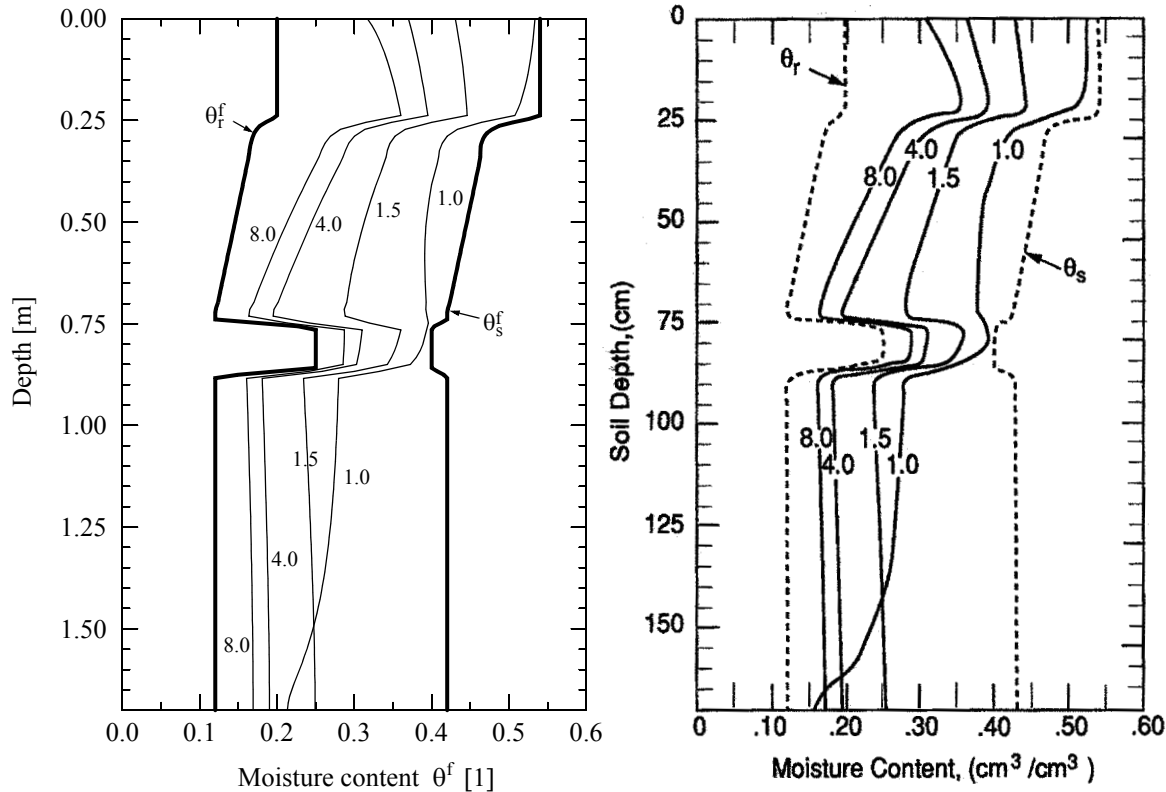

Figure 2.8 Simulated moisture-content profiles during redistribution: present solutions (left) and van Genuchten's results[15,31] (right), time in days.

A comparison between the present solutions and the results obtained by van Genuchten[15] who used a Hermitian-finite element approach is exhibited in Figs. 2.7 and 2.8. Figure 2.7 displays the simulated moisture-

content profiles during the infiltration period at $t \leq 1$ day. The moisture-content histories during the redistribution phase at $t > 1$ day of the soil column are seen in Fig. 2.8. As shown the agreement of the results is nearly perfect.

2.7.2 Drainage experiment

Vauchaud *et al.*[32] reported experimental results

which referred to a ditch-drained soil problem. Their results are useful for proving and comparing numerical schemes applied to a typical drainage problem as already done by Gureghian[33], Nützmann[34] and Nguyen[35]. A half drain-spacing with a length of 3 m and a height of 2 m is considered (Fig. 2.9).
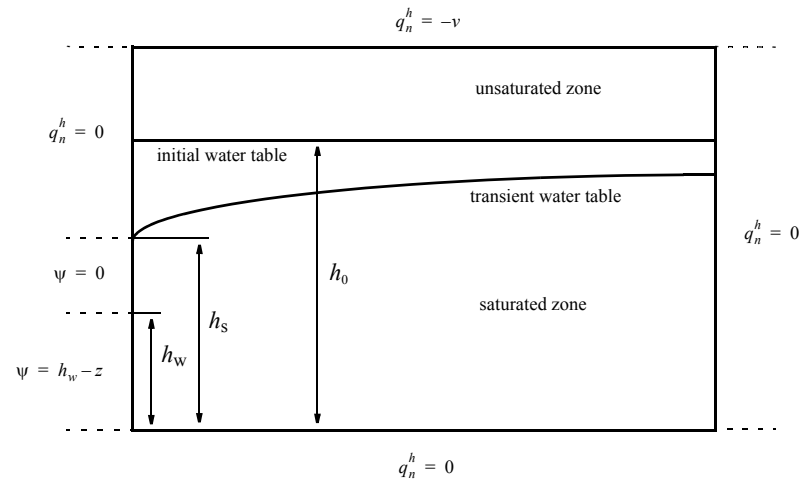

Figure 2.9 Sketch of Vauchaud *et al.*'s drainage experiment[32]: geometry and boundary conditions.

Initially, the water level in the box is at $z = h_0$ and the system is under hydrostatic equilibrium $\psi_0 = h_0 - z$. The soil is assumed to be isotropic with a saturated conductivity of $K = 1.11 \cdot 10^{-4}$ m/s. The Haverkamp parametric model (2-11), (2-12) is used for the unsaturated soil with $\alpha = 0.063396$ m, $A = 3.6 \cdot 10^{-4}$ m, $\varepsilon = 0.3$, $s_s^f = 1$ and $s_r^f = 0$. The initial head h_0 is given by 1.45 m. The water level of the ditch h_w is 0.75 m. The magnitude h_s represents the elevation of

the seepage face which is $h_s = h_0$ at $t = 0$ and have to be determined in the solution process (so as stated in Section 2.4). In Vauchaud *et al.*'s experiment the drainage process has been performed without any infiltration ($v = 0$ on top, Fig. 2.9). Accordingly, the water table descends continuously up to reaching the water level h_w of the ditch. Figure 2.10 compare the present numerical results with Vauchaud *et al.*'s experimental data. As seen the agreement is quite well.
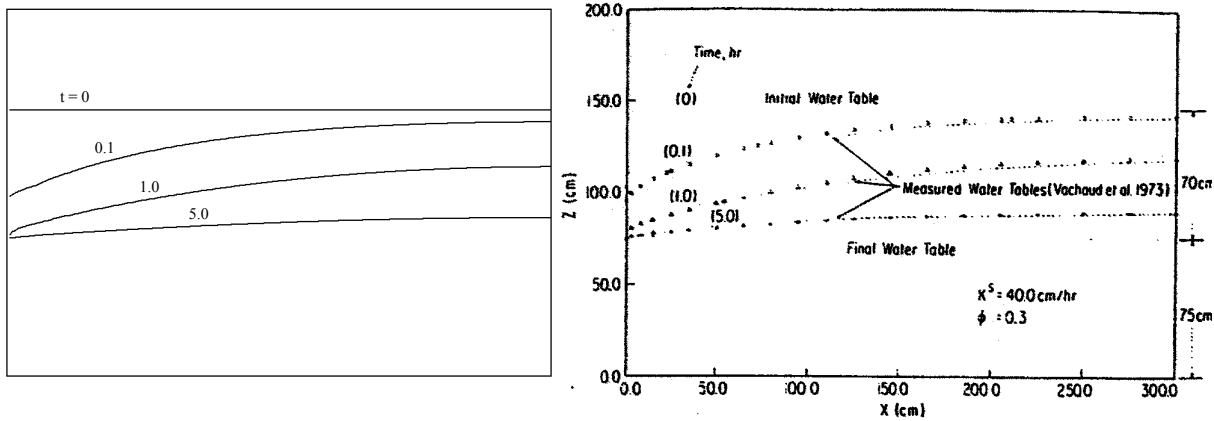
Figure 2.10 Descending water table of the drainage experiment: simulated free-surface locations (left) and water tables measured by Vauchaud *et al.*[32] (right), times in hours.
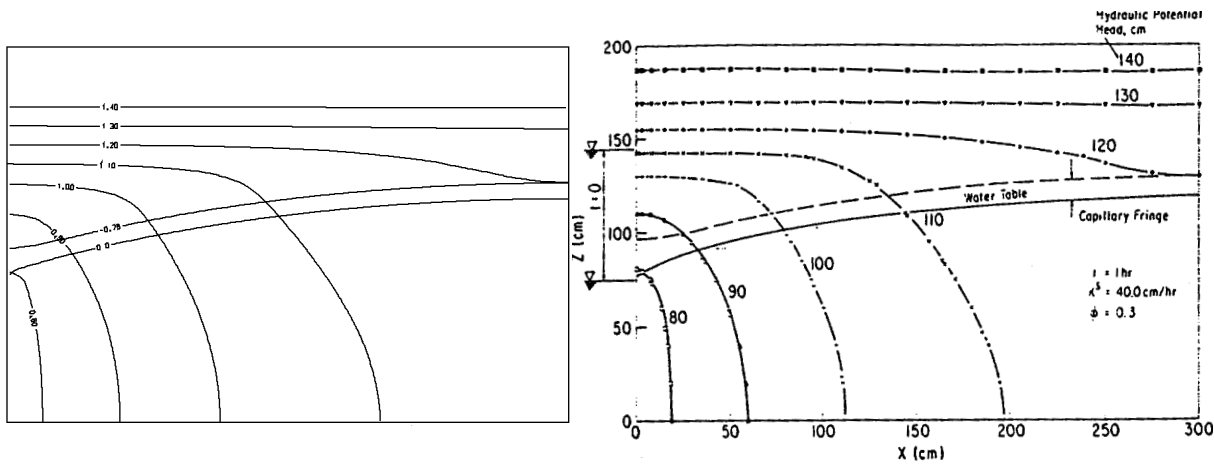

Figure 2.11 Hydraulic head contour and water table location at $t = 1$ hour: present results (left) and Gureghian's solutions[33,35] (right).

Numerical results are given by Gureghian[33]. However, its solutions are based on a non-zero infiltration rate v. A comparison of the hydraulic head contours, the water table and capillary fringe at time of $t = 1$ hour is presented in Fig. 2.11 between the present solutions and Gureghian's results. The agreement is satisfactory. Differences appear for the upper head contours which obviously results from different description of the infiltration boundary condition.

2.7.3 Dam seepage

This example is that of a transient seepage through an earth dam. The cross-sectional view and the applied finite-element mesh is displayed in Fig. 2.12. The simulations are performed to study hydrodynamic influences and effects in constructing a dam. A situation is considered where the horizontal drain of the dam fails and the drainage occurs at the dam toe and the seepage face along the slope. Figure 2.13 exhibits the simulated free surface development and the finally reached flow net if the dam consists of a homogeneous material of sand. As revealed in Fig. 2.13 the free surface reaches the dam toe where large flow gradients occur.

Expectedly, if the dam is built with a sloping core the flow regime is significantly altered. Figure 2.14 shows the simulated water table history in the dam and the equipotential and streamline pattern after reaching the final time stage. As one can see in Fig. 2.14 the zero pressure surface may intermediately have an inverted shape. It results from the antecedent moisture conditions in the unsaturated zone (redistribution of the initial moisture content). It is apparent, this situation cannot be handled with a classic free-surface modeling approach.

A 3D extension of the dam seepage problem is exhibited in Fig. 2.15 for studying the flow regime if the horizontal drain of the dam is only partially operative. It reveals how drain elements positioned at the dam toe and at a given distance can effectively dewater the downstream part of a dam consisting of homogeneous sandy material.
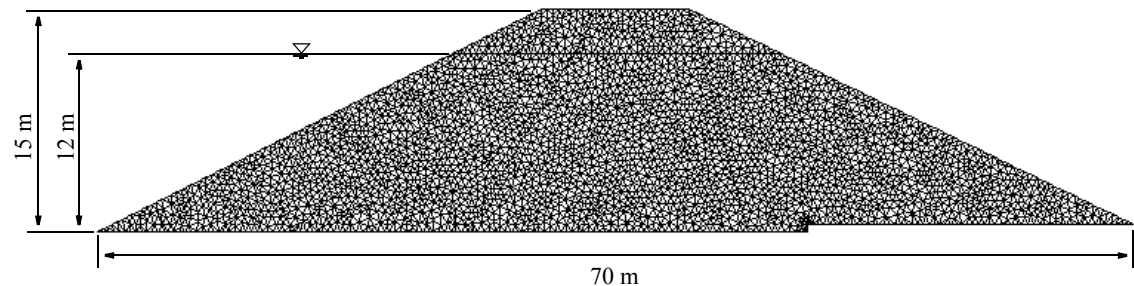
Figure 2.12 Earth dam with finite element tessellation.

a)

b)

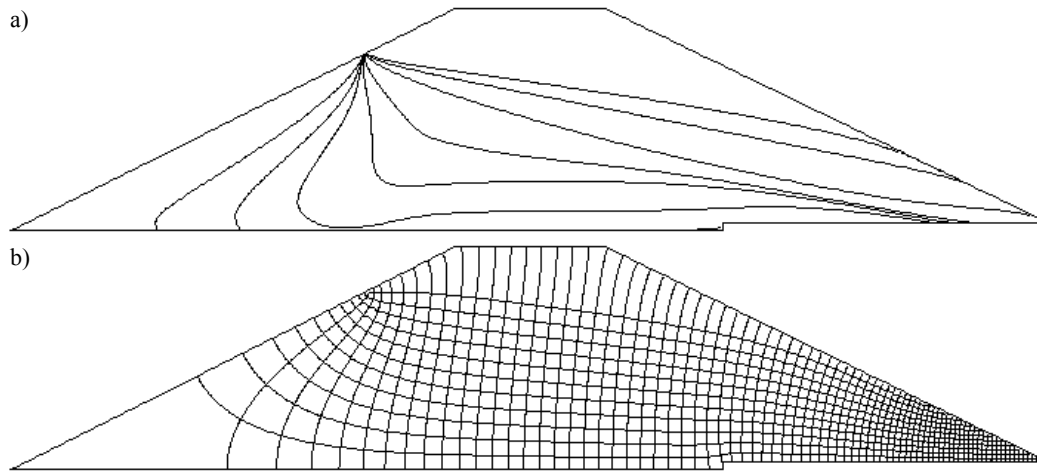**Figure 2.13** a) Advance of free surface ($\psi = 0$) in time after raising the water level in the reservoir, b) computed flow net (equipotential and streamlines) for the steady-state seepage.
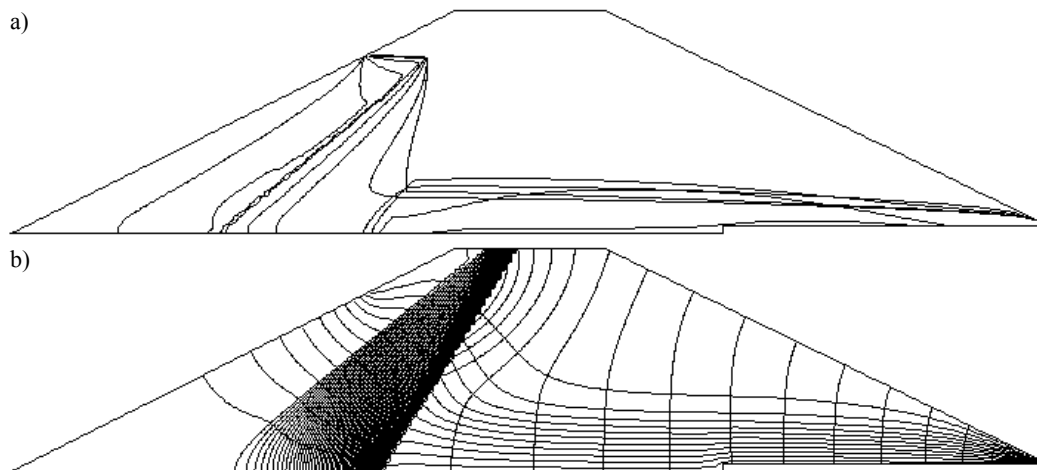
a)

b)

Figure 2.14 Solutions at a sloping core of the dam: a) advance of free surface ($\psi = 0$) in time, b) computed equipotential and streamlines for the steady-state stage.
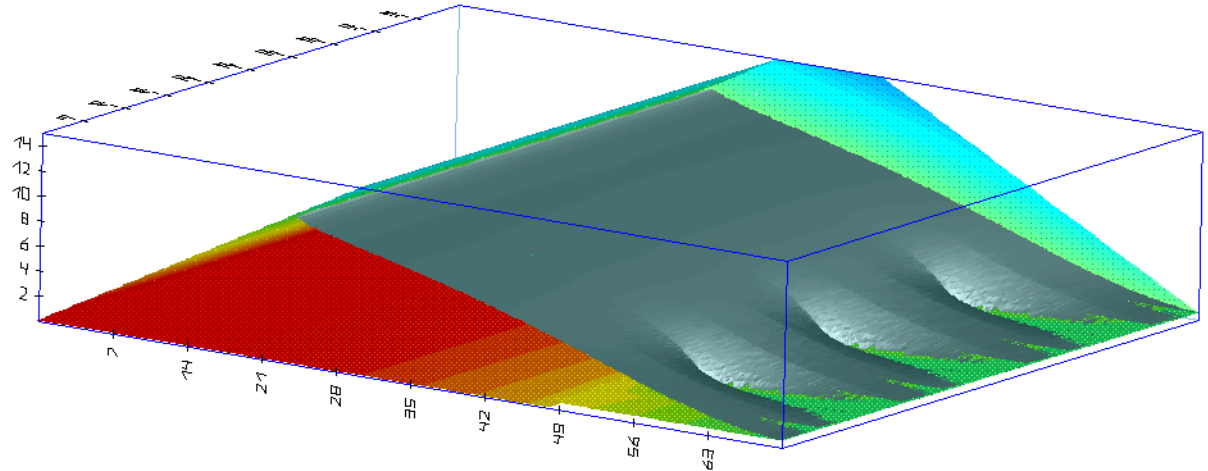
Figure 2.15 Computed 3D water table ($\psi = 0$ isosurface) for a homogeneous earth dam if partial drain elements operate.

2.7.4 Generic pit flooding test case

To compare the moving mesh strategy against fixed grid techniques of the pseudo-unsaturated modeling and the true unsaturated-saturated modeling approach let us consider a simplistic mine with a given pit geometry as shown in Fig. 2.16a. The pyramidal pit body will be filled by a water discharge of 792 m^3/d. The pit is initially dry (air-filled) and we assume for this in-pit domain a hydraulic conductivity of 100 m/s and storage coefficients of $\varepsilon_e = 1$ and $S_o = 0$. The surrounding aquifer is considered impermeable, i.e., $K = 10^{-9}$ m/s and $\varepsilon_e = S_o = 0$. As the solution the filling curve $h = h(t)$ is to be determined. For the given case an analytical solution for the filling water height can be derived

$$h = \sqrt[3]{1000 + 5.94 \cdot t} - 10 \qquad (2\text{-}40)$$

(h in meters, t in days)

Using the moving mesh only two layers are sufficient to describe exactly the stratigraphic relationships of the pit (Fig. 2.16b), where the upper layer represents the 'air' domain to be filled. On the other hand, using the pseudo-unsaturated modeling approach a fixed mesh requires more layers to adapt reasonably the slope geometry of the pit. We choose 5 layers as shown in Fig. 2.16c. As the result, the example has been proven to be a superior test case to study the accuracy of the different mesh strategies and algorithms regarding the geometry-determined temporal storage in the flooding process of the pit. As evidenced in Fig. 2.16d for the computed filling curves the moving mesh gives an excellent agreement with the analytical solutions while

the common fixed mesh strategy apparently yields quite erroneous results. The parameter alignment and

joining techniques of the BASD retain a high accuracy.
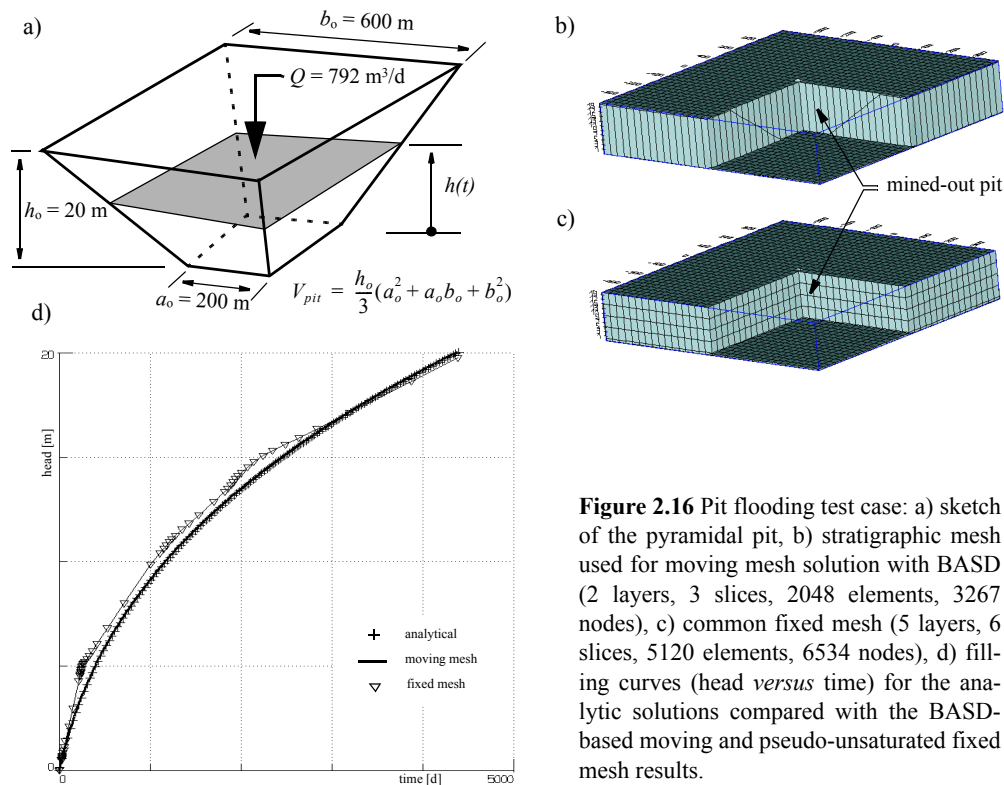
Figure 2.16 Pit flooding test case: a) sketch of the pyramidal pit, b) stratigraphic mesh used for moving mesh solution with BASD (2 layers, 3 slices, 2048 elements, 3267 nodes), c) common fixed mesh (5 layers, 6 slices, 5120 elements, 6534 nodes), d) filling curves (head *versus* time) for the analytic solutions compared with the BASD-based moving and pseudo-unsaturated fixed mesh results.

It is interesting to know whether a true unsaturated-saturated modeling approach is also appropriate to simulate even open pit flooding processes. Clearly, an air-filled mine body cannot be affected by capillary pressure relationships. Nevertheless, the variable saturation mechanism should allow to model the water table position (as the zero pressure head) in the mine regarded as

a fillable 'porous' room. The unsaturated approach serves as a contrivance to smooth the numerical solutions.

Applying the unsaturated-saturated modeling approach to the 3D pit flooding test case the van Genuchten-Mualem parametric model with $n = 2$, $A = 4$ 1/m,

$\varepsilon = 1$, $s_r^f = 1 \cdot 10^{-6}$, and $s_s^f = 1$ is assumed. The simulated filling curves are shown in Fig. 2.17. It reveals the following: The accuracy of the solution is strongly dependent on the mesh refinement. Using only 4 layers to approximate the pit body (equivalent to the above pseudo-unsaturated approach) the obtained filling curve is kinky and inaccurate. This is similar to pseudo-unsaturated approach as exhibited in Fig. 2.16. If more layers are applied the curve becomes a smooth shape and approaches to the analytical solution (Fig. 2.17).
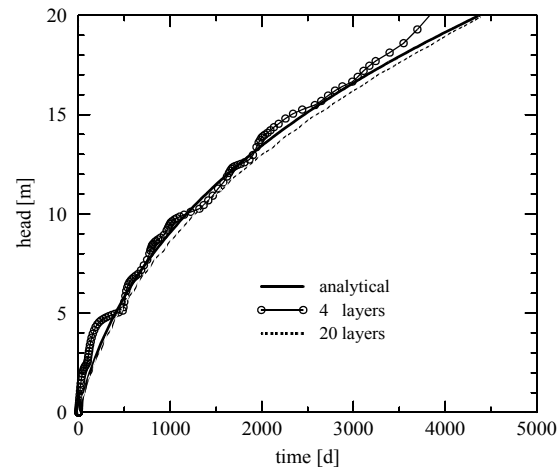
Figure 2.17 Filling curves of the 3D pit geometry simulated by the unsaturated-saturated modeling approach for different layer approximations in comparison with the analytical solution.

The mesh effect of the unsaturated-saturated modeling approach can be better shown along a very simple 2D rectangular pit geometry. Here the filling curve is simply given by $h = (Q/b) \cdot t$, where an average width of $b = 300$ m is assumed. Varying the number of ele-

ments in the vertical direction one obtains numerical filling curves such as displayed in Fig. 2.18. Apparently, the unsaturated-saturated modeling approach necessitates a sufficiently discretized approximation to find reasonable solutions for this class of problems. It becomes clear, moving mesh strategies are here superior.
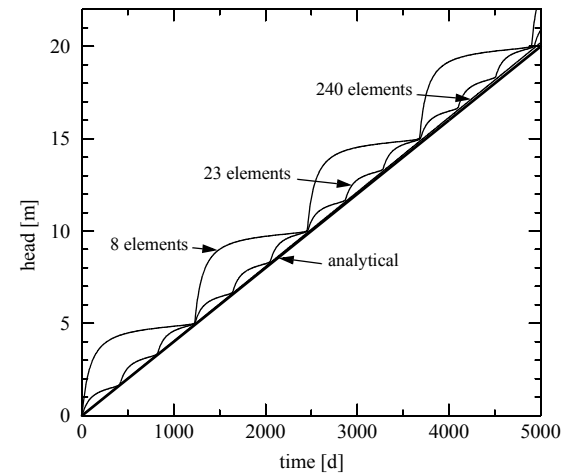
Figure 2.18 Filling curves of a 2D rectangular pit geometry simulated by the unsaturated-saturated modeling approach: mesh effects (varying number of vertical elements) in approaching to the analytical solution.

2.7.5   Real-site mining problems

To demonstrate the efficacy of the BASD-based moving mesh techniques for real-site applications two practical examples are considered. The first application refers to flooding modeling of the Königstein uranium mine[25,36]. The simulation of these pit flooding processes represents a complex task due to the complicate

hydrogeological conditions, the existence of free surfaces, density effects, high parameter contrasts, difficult geometric forms of the pit with its wide-spread network of adits, drifts, shafts, and open rooms, and the specific contaminant sources resulting from the applied in-situ uranium leaching of low-grade ores in sandstone blocks[36]. Both regional models with more than 300,000 discretized elements and local in-pit models were built up (Fig. 2.19a). The latter were used to simulate the hydraulic and contaminant transport processes in an experimental flooding area. Figure 2.19b displays the modeled hollow structure for the main stopes of the

mine incorporated into the 3D finite element mesh (Fig. 2.19a). A visual impression of the flooding behavior is given in Fig. 2.20 exhibiting the water and contaminant spreading in the pit area at two selected time stages. At early times the water primarily floods the stopes in the lower location, penetrates vertical conduits and begins to wet the blocks and 'magazines'. Contaminant matter is flushed out from the open stopes and is activated in the rocks as soon wet. At later times also the upper locations are under water and more and more blocks activate their contaminant content.
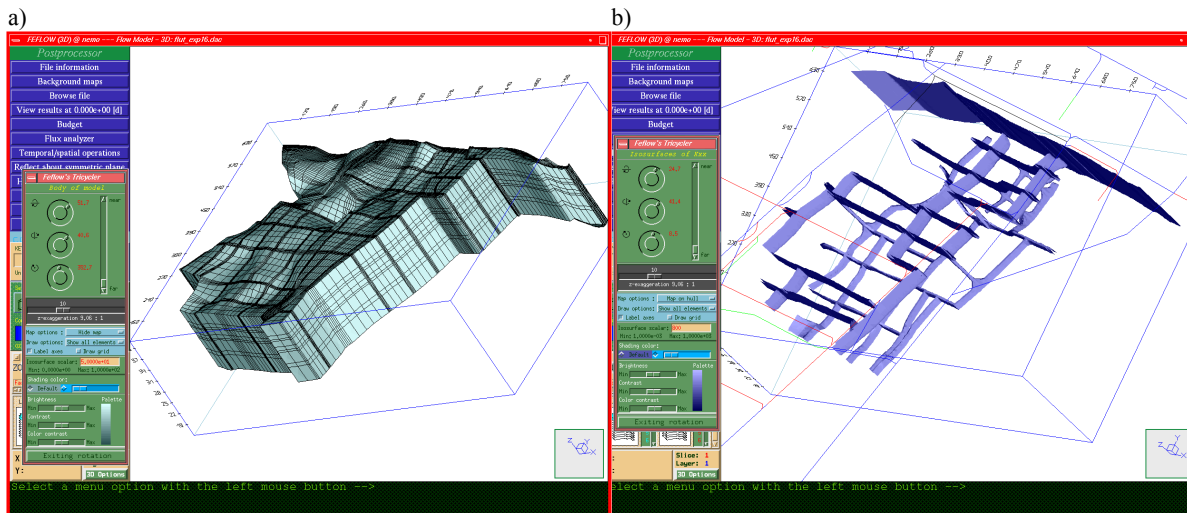
Figure 2.19 Experimental flooding area simulation of the Königstein mine: a) used 3D mesh (89,130 nodes), b) modeled main stopes by the BASD technique (isosurfaces of high conductivity).

A second example demonstrates the modeling a flooding process of a deep mine near the sea where is a threat of saltwater intrusion. Figure 2.21 exhibits the used finite element mesh for studying the water table rise and saltwater intrusion in the mine area. The

BASD technique is utilized to adapt the mesh according to the rising water table and the 3D parameter stratigraphy. Along a representative cross-section of the 3D mesh as shown in Fig. 2.22 the effect of the mesh moving becomes evidenced.
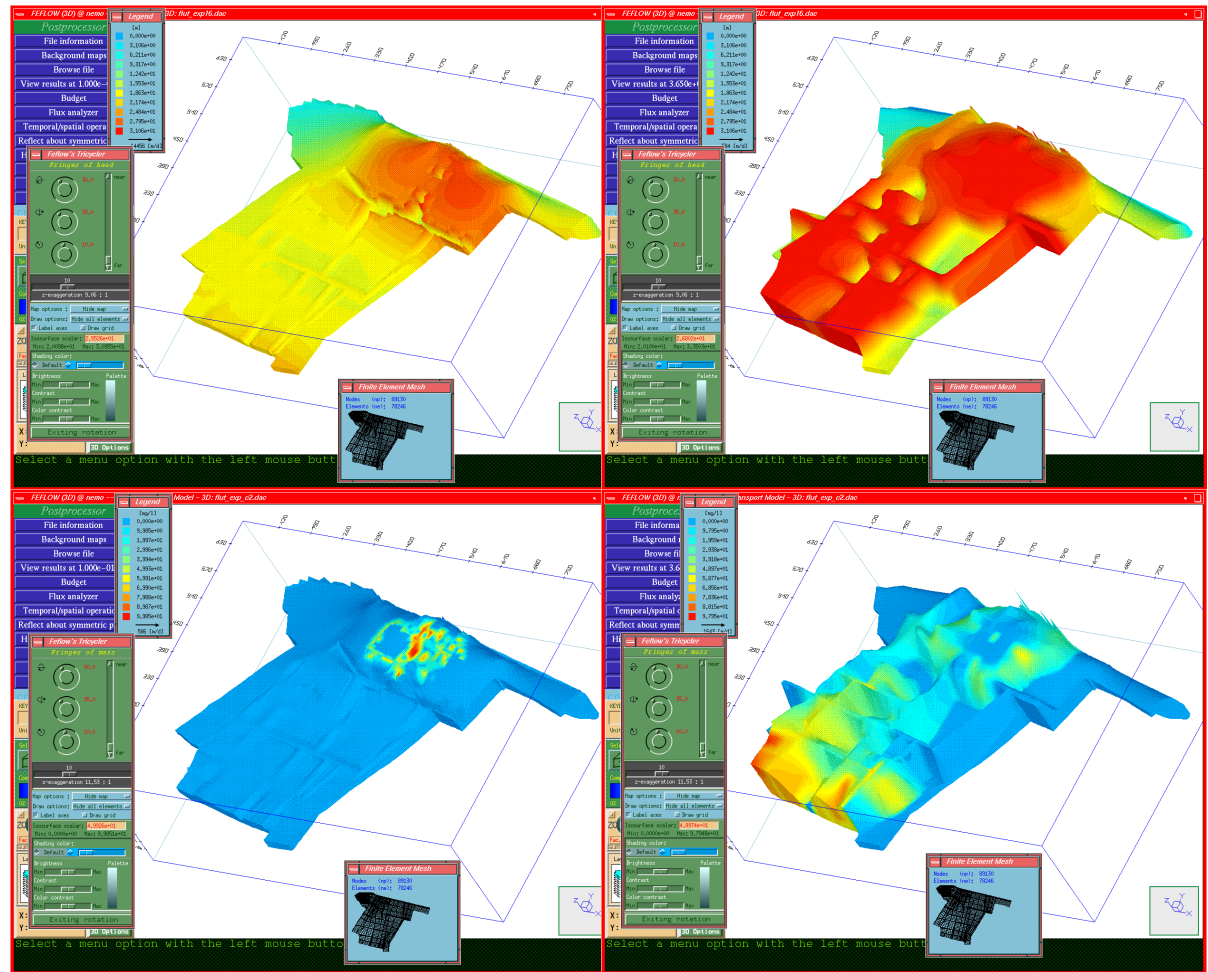
Figure 2.20 Experimental flooding area simulation of the Königstein mine: computed hydraulic head at t = 0.1 d and t = 365 d (upper), simulated contaminant distribution at t = 0.1 d and t = 365 d (lower).
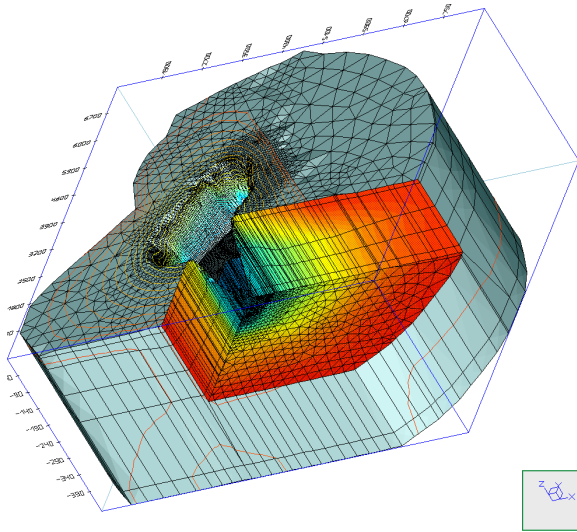
Figure 2.21 3D finite element mesh for modeling regional mine flooding and saltwater intrusion with the BASD technique.

$t = 0$

$t = 2$ yr

$t = 90$ yr
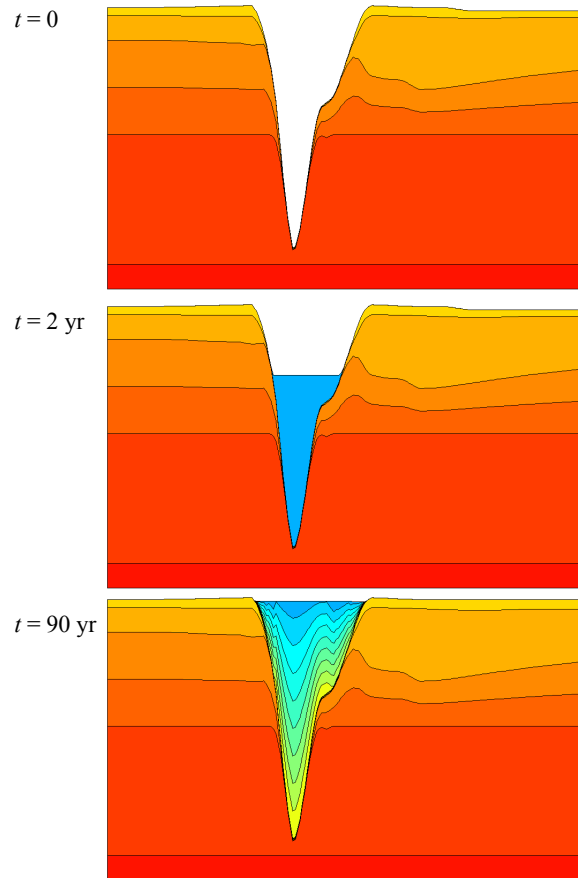
Figure 2.22 Cross-sectional view of the mine: mesh moving in adapting the rise of water table and fitting to stratigraphic properties.

2.8 Closure

Subsurface free-surface problems in 2D and 3D applications have been treated by both the unsaturated-saturated modeling approach and the groundwater free-surface modeling strategy. Fixed and moving mesh strategies are used in dependence on the modeling approaches. The finite element method is applied to solve the flow, contaminant mass and heat transport equations for variable saturation and free-surface conditions. Generalized boundary and constraint conditions are necessary to analyze free-surface problems for practical usage.

Two major approaches have been raised: (1) variably saturated models with fixed meshes, and (2) free-surface boundary condition models with adaptive (moving) meshes. A pseudo-unsaturated modeling approach as a further alternate is a physical approximation and represents neither a true physical unsaturated flow nor an exact formulation of the free-surface boundary conditions. Nevertheless, it is a widely used conception in classic groundwater models for unconfined aquifers.

A new method called BASD has been developed to adapt automatically the finite-element mesh according to the changing free surface. In this adaptation procedure all relevant data from a hydro-stratigraphic initial structure are transformed to the adapted mesh in such a manner that parameter discontinuities are maintained as best as possible (prefer parameter fit and snap before parameter interpolation and smoothing).

In a number of examples the advantages and drawbacks of the modeling approaches have been shown.

The unsaturated-saturated model represents the most general approach and is appropriate for modeling complex situations where the interaction with the vadose zone is important and multiple or non-coherent free surfaces occur. As exhibited coarse mesh approximations for unsaturated problems can lead to poor accuracy. Accordingly, an unsaturated-saturated model works nicely, but only on a sufficiently fine spatial and temporal discretization. A predictor-corrector technique with an adaptive time marching scheme of 2nd order in time has shown to be very successful in solving the highly nonlinear equations for variable saturation. In this way a h-based modeling approach gives accurate results and the resort to mixed formulations such as the $\theta - \psi$-form of the governing Richards equation is not necessary.

Since the true unsaturated-saturated modeling approach normally needs finer meshes and consequently increases the computational effort the classic free-surface fully saturated modeling approach cannot be beaten for regional groundwater problems. Here, the moving mesh strategy with the BASD technique is a powerful alternative. It has been successfully applied to complex 3D flow and transport as exemplified for mine flooding.

All described modeling approaches have been incorporated in the finite-element simulation system FEFLOW[21]. In this way, they are coming in one hand to choose the best-suited and most powerful method for the problem to be studied.

References

1. Liggett, J. A., Location of free surface in porous media. *J. Hydraul. Div., ASCE*, **103** (1977) HY4, 353-365.

2. Neuman, S. P. & Witherspoon, P. A., Finite element method of analysing steady seepage with a free surface. *Water Resour. Res.*, **6** (1970), 889-897.

3. Neuman, S. P. & Witherspoon, P. A., Analysis of nonsteady flow with a free surface using the finite element method. *Water Resour. Res.*, **7** (1971) 3, 611-623.

4. France, P. W., Finite element analysis of two- and three-dimensional unconfined seepage problems, *Proc. 1st Int. Conf. on Finite Elements in Water Resour.*, Princeton Univ., July, 1976.

5. Cheng, R. T.-S. & Li, C.-Y., On the solution of transient free surface flow problems in porous media by the finite-element method. *J. Hydrol.*, **20** (1973), 49-63.

6. Guvanasen, V. & Volker, R. E., Numerical solutions for unsteady flow in unconfined aquifers. *Int. J. Num. Methds. Engin.*, **15** (1980), 1643-1657.

7. Neuman, S. P., Saturated-unsaturated seepage by finite elements. *J. Hydraul. Div., ASCE*, **99** (1973) HY12, 2233-2250.

8. Frind, E. O. & Verge, M., Three-dimensional modeling of groundwater flow systems. *Water Resour. Res.*, **14** (1978) 5, 844-856.

9. Diersch, H.-J., Nützmann, G. & Scholz, H., Modellierung und numerische Simulation geohydrodynamischer Transportprozesse 1. Theorie. *Techn. Mechanik*, **5** (1984) 3, 44-58.

10. Diersch, H.-J. & Nützmann, G., Modellierung und numerische Simulation geohydrodynamischer Transportprozesse 2. Finite-Element-Methoden. *Techn. Mechanik*, **7** (1986) 1, 5-20.

11. Desai, C. S. & Li, G. C., Transient free surface flow through porous media using a residual procedure. *Proc. 4th Int. Symp. on Finite Element Methods in Flow Problems*, Tokyo, July, 1982, 621-632.

12. Desai, C. S., Residual flow procedure for three-dimensional free surface flow. *Proc. 7th Int. Symp. on Finite Element Methods in Flow Problems*, Huntsville, April, 1989, 1285-1296.

13. Bruch, J. C., Jr., A survey of free boundary value problems in the theory of fluid flow through porous media: variational inequality approach - Part I & II. *Adv. Water Resour.*, **3** (1980), 65-80 & 115-124.

14. Lin, C. S., Bruch, J. C., Jr., Sloss, J. M. & Comincioli, V., On the solution of transient free-surface flow problems in porous media by a fixed-domain method. *J. Hydrol.*, **7** (1984), 353-376.

15. van Genuchten, M. Th., Mass transport in saturated-unsaturated media: One-dimensional solutions. Research Rep. No. 87-WR-11, Water Resources Program, Princeton Univ., 1978.

16. Yeh, G.-T., Chen, J.-R. & Bensabat, J. A., A three-dimensional finite-element model of transient free surface flow in aquifers. In: *Computational Methods in Water Resources X*, Vol. **1**, X. Intern. Conf. on Comp. Methods in Water Resources, July 1994, Heidelberg, Kluwer Academic Publishers, 131-138.

17. Knupp, P., A moving mesh algorithm for 3-D regional groundwater flow with water table and seepage face. *Adv. Water Resour.*, **19** (1996) 2, 83-95.

18. Bear, J. *Hydraulics of groundwater.* McGraw Hill, New York, 1979.

19. Bear, J. & Verruijt, A., *Modeling groundwater flow and pollution.* Reidel Publ. Comp., Dordrecht, 1987.

20. McDonald, M. G. & Harbaugh, A. W., A modular three-dimensional finite-difference ground-water flow model. *U.S. Geological Survey.* Open-File Report 83-875, 1988.

21. Diersch, H.-J.G., Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.

22. Diersch, H.-J.G., FEFLOW-Physical basis of modeling. Reference Manual - Part I, WASY Ltd., Berlin, 2002.

23. Bear, J. & Bachmat, Y., *Introduction to modeling of transport phenomena in porous media.* Kluwer Academic Publ., Dordrecht, 1991.

24. McCord, J. T., Application of second-type boundaries in unsaturated flow modeling. *Water Resources Research,* **27** (1991) 12, 3257-3260.

25. Diersch, H.-J., Albert, H., Schreyer, J. & Richter, J., Three-dimensional modeling of flow and contaminant transport processes arising in flooding the Königstein uranium pit. In: B. Merkel *et al.* (eds): *Proc. Intern. Conf. Uranium-Mining and Hydrogeology*, Verlag Sven von Loga, Köln, 1995, 121-130.

26. Diersch, H.-J., Gründler, R., Kaden, S. & Michels, I., Toward GIS-based 3D/2D groundwater contamination modeling using FEM, In: *IX Intern. Conf. on Computational Methods in Water*

Resources, Denver CO, June 9-12, Vol.1 1992, Computational Mechanics Publications, Southampton, 749-760.

27. Celia, M. A., Bouloutas, E. T. & Zarba, R. L., A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, **26** (1990) 7, 1483-1496.

28. Ju, S.-H. & Kung, K.-J. S., Mass types, element orders and solution schemes for the Richards equation. *Computer & Geosciences*, **23** (1997) 2, 175-187.

29. Freeze, R. A., Three-dimensional transient, saturated-unsaturated flow in a ground water basin. *Water Resour. Res.*, **7** (1971) 2, 347-366.

30. Bathe, K. J. & Khosgoftaar, M. R., Finite element free surface seepage analysis without mesh iteration. *Int. J. Num. Anal. Meth. Geom.*, **3** (1979), 13-22.

31. Segol, G., *Classic groundwater simulations: proving and improving numerical models.* PTR Prentice Hall, Englewood Cliffs, New Jersey, 1994.

32. Vauchaud, G., Vauclin, M. & Khanji, D., Étude expérimentale des transgerts bidimensionels dans la zone non-saturée - Application a l'étude du drainage d'une nappe a surface libre. *Houille Blance*, **1** (1973), 65-74.

33. Gureghian, A. B., A two-dimensional finite element solution scheme for the saturated-unsaturated flow with applications to flow through ditch-drained soils. *J. Hydrol.*, **50** (1981), 333-353.

34. Nützmann, G., Eine Galerkin-finite-element-Methode zur Simulation instationärer zweidimensionaler ungesättigter und gesättigter Wasserströmungen im Boden. *Acta Hydrophysica*, **XXVIII** (1983) 1/2, 37-107.

35. Nguyen, H., WHAT-FEM - A two-dimensional finite element model for simulating water flow in variably saturated soils. Technical Note, No. I.96.125, JRC, Ispra Institute, July, 1996.

36. Diersch, H.-J. G., Voigt, R., Pagenkopf, W. & Gründler, R., Three-dimensional modeling and visualization of the Königstein uranium mine using GIS and FEM. *Geowissenschaften,* **15** (1997) 2, 52-60.

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

3

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

[a]*WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany*
[b]*Centre d'Hydrogéologie, Université de Neuchâtel, Switzerland*

ABSTRACT

The Newton method represents the numerical core of the primary variable switching technique (PVST) which has shown superior to conventional approaches in both unsaturated flow and multi-phase 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 achieved convergence in the Newton method. The present paper aims at analysing the cause of discrepancies in simulating unsaturated flows. This is done in comparison of analytical solutions which are based on exponential constitutive laws.

3.1 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. As the major advantages the PVST is (1) unconditionally mass-conservative with respect to the time step size, (2) very effective and robust for dry initial conditions, (3) a Newton-based iteration method with quadratic convergence, and (4) a general analysis method suitable for single- and multi-phase 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 achieved iteration convergence, TBFN results can be 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 along analytical solutions for the Richards equation based on exponential saturation-pressure and conductivity-pressure relationships.

3.2 Model Equations

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

$$R(s, \psi) = S_o \cdot s(\psi)\frac{\partial \psi}{\partial t} + \varepsilon \frac{\partial s(\psi)}{\partial t} \tag{3-1}$$
$$- \nabla \cdot \{K_r(\psi)\boldsymbol{K}[\nabla \psi + (1+\chi)\boldsymbol{e}]\} - Q = 0$$

which has to be solved either for ψ or s. In (3-1),

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

Constitutive relationships are additionally required (1) 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 (2) for the relative hydraulic conductivity K_r as a function of either the pressure head ψ or the saturation s:

$$\left. \begin{array}{ll} s = f(\psi) & \psi = f^{-1}(s) \\ K_r = g(\psi) = g^*(s) \end{array} \right\} \tag{3-2}$$

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

$$\frac{s - s_r}{1 - s_r} = s_e = \left\{ \begin{array}{ll} \exp[\alpha(\psi + \psi_a)] & \text{for} \quad \psi < \psi_a \\ 1 & \text{for} \quad \psi \ge \psi_a \end{array} \right. \tag{3-3}$$
$$K_r = s_e$$

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

3.3 Newton Method and PVST

The discretized form of the basic Richards equations (3-1) yields

$$\boldsymbol{R}^{n+1}(\boldsymbol{X}) = 0 \tag{3-4}$$

to be solved for a primary variable

$$\boldsymbol{X} \in (\Psi, \boldsymbol{s}) \tag{3-5}$$

which can be either c or s at the new time level $n+1$. Applying the Newton method to (3-4) we solve

$$\boldsymbol{J}^X(\psi_\tau^{n+1}, s_\tau^{n+1})\Delta \boldsymbol{X}_\tau^{n+1} = -\boldsymbol{R}_\tau^{n+1}(\Psi, \boldsymbol{s}) \tag{3-6}$$

with the increment

$$\Delta X_\tau^{n+1} = X_{\tau+1}^{n+1} - X_\tau^{n+1} \qquad (3\text{-}7)$$

and the Jacobian J^X expressed in indicial notation as

$$J_{IJ}^X(\Psi_\tau^{n+1}, s_\tau^{n+1}) = \frac{\partial R_I^{n+1}(\Psi_\tau^{n+1}, s_\tau^{n+1})}{\partial X_{\tau J}^{n+1}} \qquad (3\text{-}8)$$

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

3.4 The Nitty-Gritty

Generally, the control of the solution of the resulting highly nonlinear matrix system (3-6) 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 $\left\| d_\tau^{n+1} \right\|_{L_p}$ are the controlling criteria, which are functions of the solution differences $d_\tau^{n+1} \sim \Delta X_\tau^{n+1}$:

$$\left\| d_\tau^{n+1} \right\|_{L_p} < \delta \qquad (3\text{-}9)$$

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 (3-9) 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 ($\tau = 1$). Alternatively to the deviatory error estimate $\left\| d_\tau^{n+1} \right\|_{L_p}$, the residual $\left\| R_\tau^{n+1} \right\|_{L_p}$ may be directly controlled, such as

$$\left\| R_\tau^{n+1} \right\|_{L_p} < \delta_2 \left\| F^{n+1} \right\|_{L_p} \qquad (3\text{-}10)$$

where an additional error tolerance δ_2 appears and an appropriate normalisation of the residual (here with respect to the external supply F^{n+1}) is required. In the TBFN deviatory errors $\left\| d_\tau^{n+1} \right\|_{L_p}$ and residual errors $\left\| R_\tau^{n+1} \right\|_{L_p}$ can be alternatively employed. Instead of a one-step Newton control so as done in the PCOSN the predictor-corrector technique can also be extended to a

multiple step Newton (PCMSN) strategy satisfying both criteria (3-9) and (3-10). To measure the global balance error we introduce

$$\Re(T) = \frac{\displaystyle\int_{t=0}^{T} \left\| \boldsymbol{R}_\tau(t) \right\|_{L_2} dt}{\displaystyle\int_{t=0}^{T} \left\| \boldsymbol{F}(t) \right\|_{L_2} dt} \qquad (3\text{-}11)$$

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

3.5   Analytical Solution

In one dimension the Richards equation (3-1)

$$\varepsilon \frac{\partial s(\psi)}{\partial t} - \frac{\partial}{\partial z}\left[K_r(\psi) K\left(\frac{\partial \psi}{\partial z} - 1 \right) \right] = 0 \qquad (3\text{-}12)$$

can be transformed into the linear advective-dispersive equation of the form

$$\frac{\partial s}{\partial t} + \frac{K}{\varepsilon(1-s_r)}\frac{\partial s}{\partial z} - \frac{K}{\varepsilon\alpha(1-s_r)}\frac{\partial^2 s}{\partial z^2} = 0 \qquad (3\text{-}13)$$

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

With $s(z, 0) = s_i$ and $s(0, t) = s_o$ the solution is

$$s(z, t) = s_i + \frac{(s_o - s_i)}{2}\left\{ \operatorname{erfc}\left(\frac{z - \dfrac{Kt}{\varepsilon(1-s_r)}}{2\sqrt{\dfrac{Kt}{\varepsilon\alpha(1-s_r)}}} \right) \right.$$
$$\left. + e^{\alpha z}\operatorname{erfc}\left(\frac{z + \dfrac{Kt}{\varepsilon(1-s_r)}}{2\sqrt{\dfrac{Kt}{\varepsilon\alpha(1-s_r)}}} \right) \right\} \qquad (3\text{-}14)$$

It can be easily seen from (3-13) that with large α the problem is dominated by advection. Otherwise, considering a fully implicit time discretization the temporal numerical dispersion can be estimated to

$$D_{\text{numdisp}}^{n+1} \approx \frac{\Delta t_n}{2}\left[\frac{K}{\varepsilon(1-s_r)} \right]^2 \qquad (3\text{-}15)$$

3.6   Test Case

The problem is described in Fig. 3.1. For the lower boundary a free drainage-type boundary condition is applied (Diersch, 1998). The 6 m column is discretized by 120 linear elements, so the nodal spacing becomes $\Delta z = 5$ cm.

3.6.1   Newton control by the deviatory error criterion (3-9)

The computed saturation profiles for two α-parameters in comparison with the analytical solution are

shown in Fig. 3.2. 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 $\Delta t_{max} = 0.2 \text{ d}$ and $\Delta t_{max} = 0.05 \text{ d}$, respectively. It leads to the total integral balance errors (3-11) of $\Re(3 \text{ d}) \approx 160\%$ and $\approx 40\%$, respectively. In contrast, the PCOSN took 240 variable time and Newton steps resulting an acceptable balance error of only $\Re(3 \text{ d}) \approx 0.06\%$ with a very good agreement with the analytical solution. Similar results appear for $\alpha = 200 \text{ m}^{-1}$, where the TBF gives $\Re(3 \text{ d}) \approx 7\%$ (74 time and 344 Newton steps) and $\Re(3 \text{ d}) \approx 2\%$ (107 time and 301 Newton steps), respectively, while the PCOSN obtains $\Re(3 \text{ d}) \approx 0.09\%$ after 360 time and Newton steps.

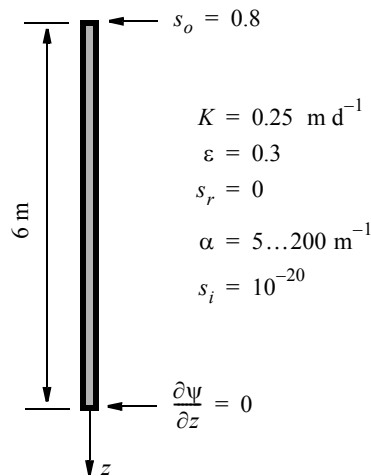## 3.6.2 Newton control by the residual error criterion (3-10)

As outlined in Fig. 3.3 the conservative problems disappears for the TBFN if the residual error criterion (3-10) is used with $\delta_2 = 10^{-4}$. The adaptive PCMSN and the TBFN gives comparable results which agree quite well with the analytical solution. Here, the PCMSN is now controlled by two the criteria (3-9) and (3-10): $\delta = 10^{-4}$ 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 $\Re(3 \text{ d}) \approx 0.0006\%$ and $\approx 0.001\%$, respectively. Expectedly, the PCMSN took more time steps. We found 164 time (268 Newton) steps with $\Re(3 \text{ d}) \approx 0.01\%$ and 165 time (362 Newton) with $\Re(3 \text{ d}) \approx 0.001\%$ for $\alpha = 5 \text{ m}^{-1}$ and $\alpha = 200 \text{ m}^{-1}$, respectively.

$$s_o = 0.8$$

$$K = 0.25 \text{ m d}^{-1}$$
$$\varepsilon = 0.3$$
$$s_r = 0$$
$$\alpha = 5...200 \text{ m}^{-1}$$
$$s_i = 10^{-20}$$

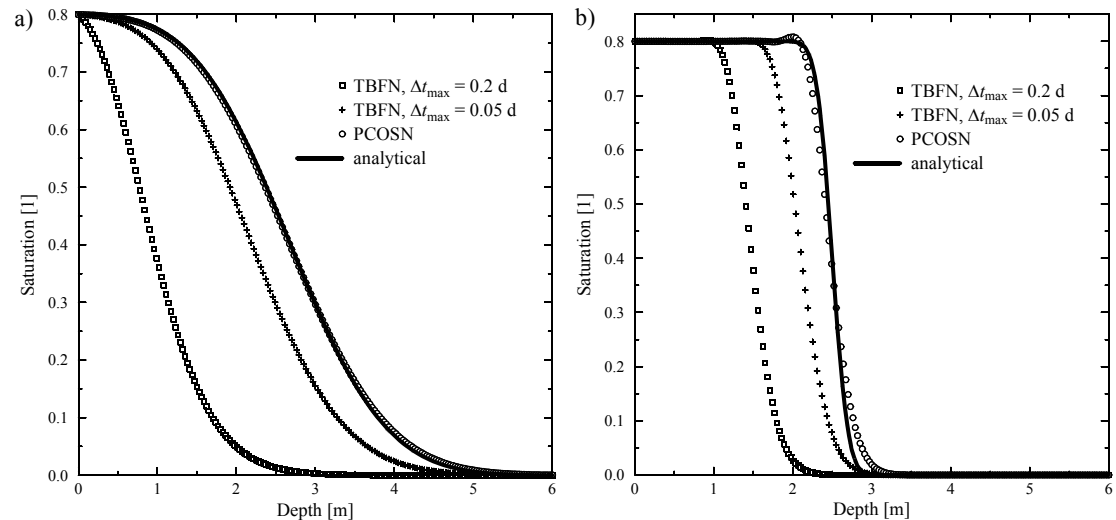$$\frac{\partial \psi}{\partial z} = 0$$

Figure 3.1 Sketch of the test problem.

Figure 3.2 Saturation profiles at $t = 3$ d for a) $\alpha = 5$ m^{-1} and b) $\alpha = 200$ m^{-1}, $\delta = 10^{-4}$ with the deviatory error criterion (3-9) for the Newton control, aggressive target change parameters are used for the TBFN with a maximum time step constraint Δt_{max}.

3.7 Conclusions

For the TBFN the residual error criterion should be preferred rather than standard deviatory tests to avoid 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, so as done in the PCMSN scheme, is not necessary in the most cases.

References

1. Diersch, H.-J. G. (1998) Treatment of free surfaces in 2D and 3D groundwater modeling. *Math. Geologie* **2,** 17-43.
2. Diersch, H.-J. G. & Perrochet, P. (1999) On the primary variable switching technique for simulating unsaturated-saturated flows. *Adv. Water Resour.* **23** (1)**,** 25-55.
3. Forsyth, P. A., Wu, Y. S. & Pruess, K. (1995) Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. *Adv. Water Resour.* **18**(1)**,** 25-38.
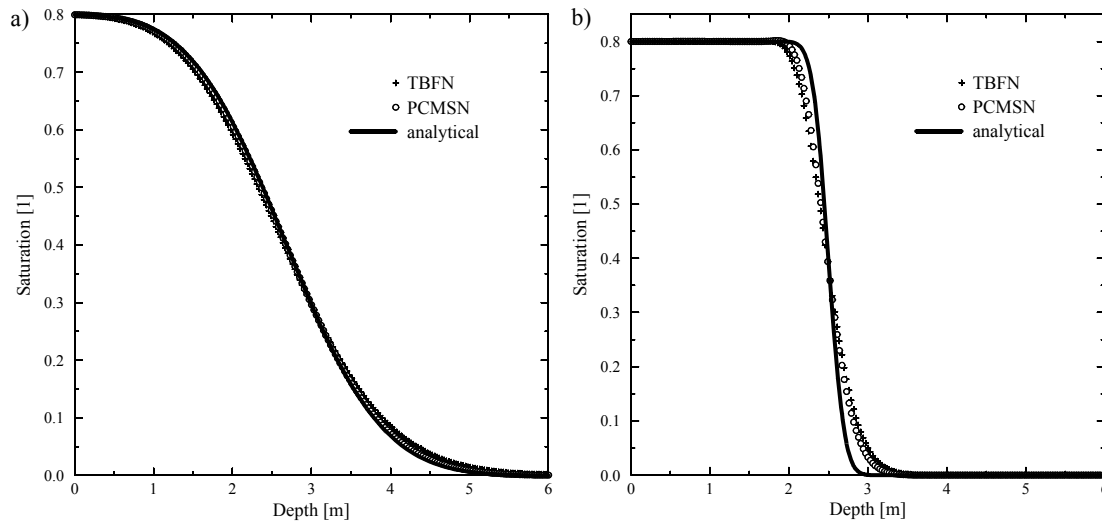
Figure 3.3 Saturation profiles at $t = 3$ d for a) $\alpha = 5 \ \mathrm{m}^{-1}$ and b) $\alpha = 200 \ \mathrm{m}^{-1}$, $\delta = \delta_2 = 10^{-4}$ with the residual error criterion (3-10) for the Newton control.

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

A shock-capturing finite-element technique for unsaturated-saturated flow and transport problems

4

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

ABSTRACT

A shock-capturing technique is introduced for a finite element approach of the governing balance equations of flow, contaminant mass and heat transport in variably saturated porous media. It represents a nonlinear method which depends itself on the numerical solution and is characterized by an additional discontinuity-capturing term controlling the derivatives in the direction of the solution gradient. The technique is embedded in a predictor-corrector scheme of first and second order in time to handle the solution implicitly. The impact of the shock-capturing technique is studied for selected applications.

4.1   Introduction

In the numerical modeling of multidimensional transport problems upwind techniques such as SUPG (streamline-upwind Petrov-Galerkin)[1] or scalar upstream weighting[2] are standard to stabilize the solutions when convection becomes highly dominant. While the classic artificial diffusion method often suffers in a considerable smearing of steep fronts[3] the SUPG formulation cannot preclude the presence of overshoots and undershoots in the vicinity of sharp gradients[4]. For nonlinear situations, e. g., buoyancy-driven convection, such type of oscillations may affect the global stability of the numerical results and the solutions fail.

It has been shown[4] the streamline is not always the appropriate upwind direction. A generalization of the streamline concept in form of adding an additional discontinuity-capturing term was presented by Hughes and Mallet[5]. The shock capturing (SC) method applied to finite elements has been developed by Johnson *et al.*[6] and Codina[7]. The SC technique appears as a nonlinear method, that is, the scheme depends itself on the numerical solution. The main idea behind shock capturing is to increase the amount of damping in the neighborhood of layers. Then, the damping to be added must be proportional to the element residual and must be vanish in regions where the solution is smooth and also where the convective term of the residual is small.

In this work the SC technique is applied to multidimensional nonlinear transport processes arising in density-driven and unsaturated porous medium flows.

4. A shock-capturing finite-element technique for unsaturated-saturated flow and transport problems

4.2 Shock Capturing Technique

The basic idea is to employ an asymmetric weak formulation

$$\tilde{w} = w + \tau_1 v \cdot \nabla w + \tau_2 v_\| \cdot \nabla w \qquad (4\text{-}1)$$

where the first term is the standard Galerkin test function, the second term is the linear SUPG modification and the third term is the nonlinear discontinuity-capturing operator. The vector $v_\|$ is the projection of the velocity vector v onto the direction of the local gradient $\nabla \psi$ of a solution ψ. Then, the parameters τ_1 and τ_2 are defined as

$$\tau_1 = \frac{\alpha h}{2|v|} \qquad \tau_2 = \max\left(0, \frac{\alpha h}{2|v_\||} - \tau_1\right) \qquad (4\text{-}2)$$

$$v_\| = \frac{v \cdot \nabla \psi}{|\nabla \psi|^2} \nabla \psi$$

where α is an upwind parameter[2] and h is the element size. Considering, for simplicity, the transient convection-diffusion transport equation of the scalar quantity ψ

$$\dot{\psi} + v \cdot \nabla \psi - \nabla \cdot (D \cdot \nabla \psi) - Q = 0 \qquad (4\text{-}3)$$

where D is the tensor of hydrodynamic dispersion and Q is a sink/source, and introducing the spatial residual

$$R(\psi) = v \cdot \nabla \psi - \nabla \cdot (D \cdot \nabla \psi) - Q \qquad (4\text{-}4)$$

the weak form of (4-3) by using the generalized test function (4-1) introduces an isotropic shock capturing dispersion coefficient in the form[6,7]

$$D_{sc} = \frac{1}{2}\alpha_c h \frac{|R(\psi)|}{|\nabla \psi|} \qquad (4\text{-}5)$$

if $|\nabla \psi| \neq 0$ and zero otherwise. The upwind function α_c is given by

$$\alpha_c = \max\{0, a - 1/\gamma_\|\} \qquad \gamma_\| = \frac{|v_\||h}{2D_d} \qquad a \approx 0.7 \qquad (4\text{-}6)$$

In principle, the shock capturing dispersion coefficient D_{sc} could be added to the dispersion tensor both in the longitudinal and transverse direction of the flow. However, a lower order time discretization can already introduce a damping measure along the streamlines of magnitude

$$D_{\Delta t} = \frac{\Delta t}{2}|v|^2 \qquad (4\text{-}7)$$

and the final shock capturing is anisotropic and the dispersion tensor yields the form

$$D = (\tilde{\beta}_L - \tilde{\beta}_T)\frac{v \otimes v}{|v|} + (\tilde{\beta}_T|v| + D_d)I \qquad (4\text{-}8)$$

with the total longitudinal and transverse dispersivities, respectively,

$$\left.\begin{array}{l} \tilde{\beta}_L = \beta_L + \dfrac{\max\{0, D_{sc} - D_{\Delta t}\}}{|v|} \\[4mm] \tilde{\beta}_T = \beta_T + \dfrac{D_{sc}}{|v|} \end{array}\right\} \qquad (4\text{-}9)$$

where β_L, β_T are the true 'physical' dispersivities, D_d is the diffusion coefficient, and I is the unit tensor. Notice, for a higher-order time discretization (e.g., Crank-Nicolson scheme) $D_{\Delta t}$ is to be neglected.

The terms corresponding to the shock capturing dispersion (4-5) are nonlinear and an appropriate numerical treatment is required. SC techniques are often employed with explicit time discretization. However, due to the strong stability bounds of explicit approaches for the solution of the present problems an implicit version of the SC technique is preferred. It is based on a predictor-corrector scheme which is implemented for unsaturated-saturated flow, mass and heat transport processes.

4.3 Implementation

The 2D and 3D finite-element discretization of the coupled unsaturated-saturated flow, mass and heat transport problems leads to the following matrix system

$$O(S)\dot{h} + M\dot{S} + K(h, S)h = F(h, S, C, T)$$
$$Aq = B(h, S, C, T)$$
$$P(S, C)\dot{C} + D(q, h, S, C, T)C = R(S, C)$$
$$U(S)\dot{T} + L(q, h, S, C, T)T = W(S, T)$$

(4-10)

to solve the hydraulic head h, the fluid saturation S, the Darcy fluxes q, the concentration C, and the temperature T. A first order FE/BE (Forward Euler/Backward Euler) and a second order AB/TR (Adams-Bash-

forth/Trapezoid Rule) predictor-corrector approach with a one-step full Newton technique are applied[8]. Finally, the semi-implicit corrector step yields the following matrix form:

$$\left[J_h + \left(\frac{\varphi O}{\Delta t_n} + K \right) + \Theta \frac{\varphi M}{\Delta t_n} \right] h^{n+1} =$$

$$= \left[J_h + \Theta \frac{\varphi M}{\Delta t_n} \right] h_p^{n+1} + O\left(\frac{\varphi}{\Delta t_n} h^n + \kappa \dot{h}^n \right) -$$

$$-M\left[\frac{\varphi}{\Delta t_n}(S_p^{n+1} - S^n) - \kappa \dot{S}^n \right] + F^{n+1}$$

(4-11)

$$Aq^{n+1} = B^{n+1}$$

$$\left[J_c + \left(\frac{\varphi P}{\Delta t_n} + D \right) \right] C^{n+1} = J_c C_p^{n+1} + P\left(\frac{\varphi}{\Delta t_n} C^n + \kappa \dot{C}^n \right) + R^{n+1}$$

$$\left[J_t + \left(\frac{\varphi U}{\Delta t_n} + L \right) \right] T^{n+1} = J_t T_p^{n+1} + U\left(\frac{\varphi}{\Delta t_n} T^n + \kappa \dot{T}^n \right) + W^{n+1}$$

where φ is 2 for the TR and 1 for the BE scheme, κ is 1 for the TR and 0 for the BE scheme, $\Theta = ds/dh|_p^{n+1}$ is the derivation of the saturation with respect to the hydraulic head and is a given function, n designates the time level, subscript p indicates the explicit predictor solutions obtained from FE or AB schemes[8], and J represents the corresponding partial Jacobian of the Newton method. The nonlinearities appearing in the matrices and which are caused by both the physical problem and the nonlinear shock-capturing parameter (4-5) are linearized by using the predictor values. The solution is automatically controlled by an error-based timestep adaptation. Notice, as seen in (4-11) the unsaturated flow equations of the Richards type are treated in a mixed saturated-head variable formulation possessing good conservative properties.

4.4 Numerical Results

4.4.1 Hoopes and Harlemann's two well problem

Hoopes and Harlemann[9] performed a number of experiments using a sand model to measure the distribution of a solute between recharging and discharging wells in the saturated zone. Figure 4.1a shows the coarse unstructured 2D mesh used to compare the results at time $t = 0.2$ d for the Galerkin method (Fig 4.1b), the SUPG scheme (Fig. 4.1c) and the proposed

SC method (Fig. 4.1d). Following parameters have been used: injection rate $q = 6.433$ m^2/d, $D_d = 0$, $\beta_L = 0.0015$ m, $\beta_T = 0$ (for more details see[10]).

It reveals the Galerkin scheme produces significant wiggles while the SUPG and the SC schemes give stable solutions. The SC results appear superior to the SUPG distribution. The SC scheme introduces a smaller amount of upwind and exhibits more accurate results.
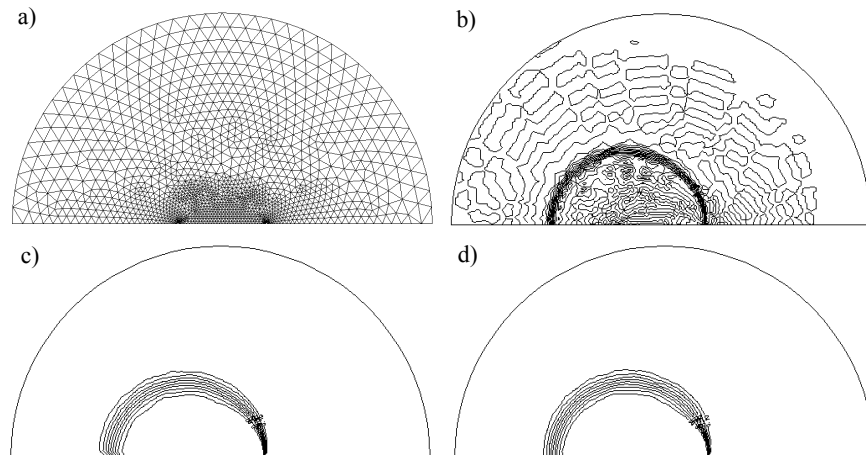
Figure 4.1 Comparison for Hoopes and Harlemann's experiment: a) used mesh, b) Galerkin (no upwind), c) SUPG, d) shock capturing.

4.4.2 Sinking of a heavy solute in a 'sealed box'

The advantage of the SC method becomes more apparent if considering the following density-driven problem (Fig 4.2).

a)

b)
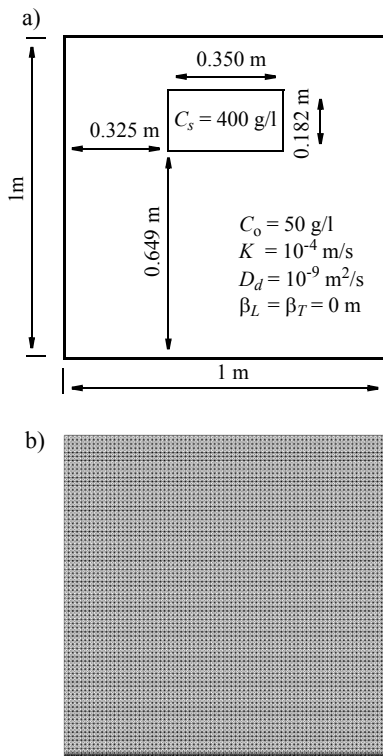
Figure 4.2 a) domain, b) used mesh (17,336 nodes; 33,198 triangles).

a)

b)

c)

d)
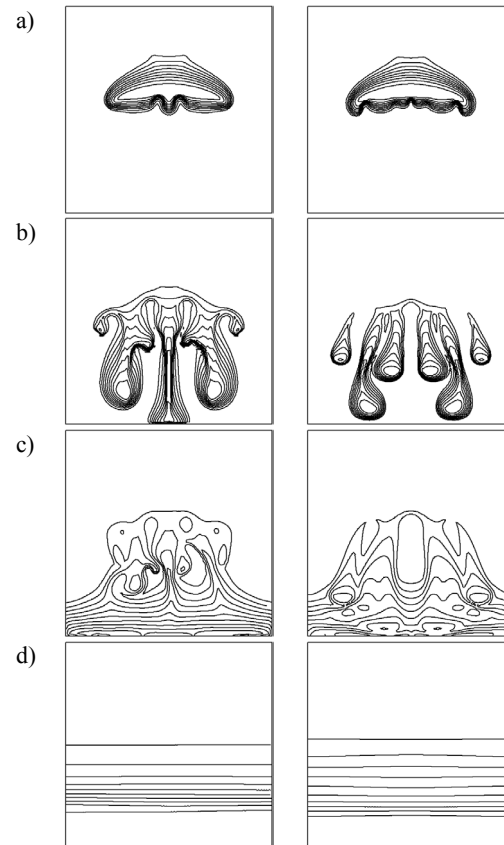
Figure 4.3 Computed solute distribution for the SUPG (left column) and the SC scheme (right column) at different times: a) 0.5 d, b) 1.5 d, c) 2.5 d, and d) 100 d.

4. A shock-capturing finite-element technique for unsaturated-saturated flow and transport problems

In a 2D box of a saturated porous medium with impervious boundaries a heavy solute with a concentration of 400 g/l is initially placed at an upper location. The background concentration is 50 g/l and mechanical dispersion should be neglected. It becomes clear that the gravity-induced sinking process of the solute is difficult to model because the buoyancy forces are very large. Otherwise, it is to be expected that discretization effects becomes significant due to different damping measures required in the numerical solution. An unstructured mesh of a medium resolution is used for comparing the SUPG and SC schemes. The results are displayed in Fig. 4.3. As shown the evolving pattern formations are quite different for both schemes. It clearly reveals the overdiffusive property of the SUPG scheme against the superior SC method. Notice, a Galerkin approach cannot be successful for the used mesh.

4.4.3 Infiltration into an initially dry region

This unsaturated heterogeneous 2D problem was introduced by Forsyth and Kropinski[11] to study upstream weighting methods. The problem is difficult to solve due to both the initially dry condition (initial capillary pressure head of -100 m) and the flat capillary pressure curve (van Genuchten pore size distribution index of 5). The flow problem[11] and an accompanying transport problem of an infiltrated tracer (with $D_d = 10^{-9}$ m^2/s, $\beta_L = 0.05$ m, $\beta_T = 0.005$ m) is simulated by using the SC method on a coarse 90x21 quadrilateral mesh, which is comparable to[11]. The resulting distributions for the saturation and the solute at time $t = 30$ d

are exhibited in Fig. 4.4. A comparison with Forsyth and Kropinski's saturation patterns indicates that the present results possess a good adaptation of steep gradients without spurious oscillations. Such a steepness was achieved by Forsyth and Kropinski only at a finer mesh. There, the agreement is quite well.
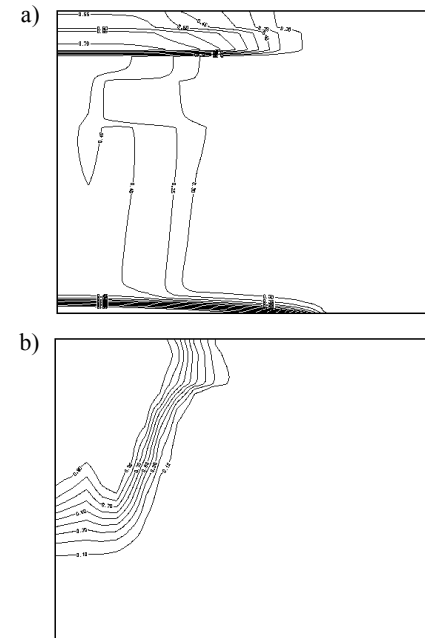
Figure 4.4 Simulated distributions of a) saturation and b) solute for the Forsyth and Kropinski's problem[11] at $t = 30$ d by using the SC method for a coarse 90x21 quadrilateral finite element mesh.

4.5 Conclusions

The nonlinear SC method have shown to be superior to a common SUPG scheme. It improves numerical stability by introducing crosswind dispersion in the neighborhood of layers. It is less overdiffusive and does not require too much computational extra cost. Embodied into a predictor-corrector technique the SC method can be used in an implicit mode and is easy to implement. This makes it an attractive technique for practical usage. The SC method is available in the FEFLOW code[10].

References

1. Brooks, A.N. & Hughes, T.J.R., Streamline upwind Petrov-Galerkin formulations for convection-dominated flows with particular emphasis on the incompressible Navier-Stokes equations, *Comput. Meths. Appl. Mech. Engrg.* **32**, pp. 199-259, 1982.
2. Heinrich, J.C., Huyakorn, P.S., Zienkiewicz, O.C. & Mitchell, A.R., An upwind finite element scheme for two-dimensional convective transport equation, *Internat. J. Numer. Meths. Engrg.* **11**, pp. 134-143, 1977.
3. Roache, P.J., On artificial viscosity, *J. Comput. Phys.* **10**, pp. 169-184, 1972.
4. Hughes, T.J.R., Mallet, M. & Mizukami, A., A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, *Comput. Meths. Appl. Mech. Engrg.* **54**, pp. 341-355, 1986.
5. Hughes, T.J.R. & Mallet, M., A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems, *Comput. Meths. Appl. Mech. Engrg.* **58**, pp. 329-336, 1986.
6. Johnson, C., Szepessy, A. & Hansbo, P., On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws, *Math. Comput.* **54**, pp. 82-107, 1990.
7. Codina, R., A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation, *Comput. Meths. Appl. Mech. Engrg.* **110**, pp. 325-342, 1993.
8. Diersch, H.-J.G. & Kolditz, O., Coupled groundwater flow and transport: 2. Thermohaline and 3D convection systems, *Adv. in Water Resour.* **21**, No. 5, pp. 401-425, 1998.
9. Hoopes, J.A. & Harlemann, D.R.F., Wastewater recharge and dispersion in porous media, *ASCE J. Hydr. Div.* **93** (HY5), pp. 51-71, 1967.
10. Diersch, H.-J.G., Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.
11. Forsyth, P.S. & Kropinski, M.C., Monotonicity considerations for saturated-unsaturated subsurface flow, *SIAM J. Sci. Comput.* **18**, pp. 1328-1354, 1997.

4. A shock-capturing finite-element technique for unsaturated-saturated flow and transport problems

Error norms used in FEFLOW

5

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

5.1 Errors and Measures

Mathematically, an error e can be considered as the difference between an *approximate* $\hat{\psi}$ and an *exact* solution ψ

$$
\begin{aligned}
e &= \psi - \hat{\psi} \\
e_i &= \psi_i - \hat{\psi}_i
\end{aligned}
\qquad (5\text{-}1)
$$

Note, ψ stands for a primary solution variable, which is typically in FEFLOW the hydraulic head h, the concentration C or the temperature T.

The specification of a local error in the form (5-1) is generally not appropriate and more general error measures are required. For this reason various 'norms' representing some integral scalar quantity are often introduced to measure errors. In an abstract sense the vector (or Hölder) norm, so-called L_p norm, is the most general expression for an error measure, viz.,

$$
\|e\|_p = \left(\sum_{i=1}^{N} |e_i|^p \right)^{1/p} \qquad p \geq 1 \qquad (5\text{-}2)
$$

where N represents the number of vector elements. In practical applications the error norm should be appropriately chosen to focus on the particular quantity of interest. As the result, in varying p one can emphasize the maximum differences occurred in the solution or one has more interest in an average and integral measure of the error. While the former is a much stronger measure and focus on local effects, the latter gives often a representative measure of the overall error in the entire solution space. Accordingly, FEFLOW uses different norms in measuring errors which can be optionally chosen.

5.2 Errors Derived for Numerical Schemes

The used specific form of the error measure depends on the numerical context. For instance, for a simple *iterative scheme* applied at a given time level it

has the form

$$|e_i| = \left|\psi_i^{\tau+1} - \psi_i^{\tau}\right| \qquad (5\text{-}3)$$

where τ corresponds to the iteration counter. On the other hand, in using the predictor-corrector techniques with automatic time stepping the errors are derived for the *Forward Euler/Backward Euler (FE/BE) scheme* as

$$|e_i| = \frac{1}{2}\left|\psi_i - \psi_i^{\Delta}\right| \qquad (5\text{-}4)$$

and for the *Adams-Bashforth/Trapezoid rule (AB/TR) scheme* as

$$|e_i| = \frac{\left|\psi_i - \psi_i^{\Delta}\right|}{3\left(1 + \dfrac{\Delta t_{n-1}}{\Delta t_n}\right)} \qquad (5\text{-}5)$$

where the superscript Δ indicates the predictor solution, the subscript n is the time level, and Δt is the time increment.

5.2.1 Euclidean L_2 integral Root Mean Square (RMS) error norm

This is the default error norm to measure an integral error quantity. It represents an error measure which is natural to the solution approximation of the governing balance equations in then sense of square-integrable functions over the solution domain:

$$\|e\|_{L_2} = \left[\frac{1}{N}\left(\frac{1}{\psi_{max}^2}\sum_{i=1}^{N}|e_i|^2\right)\right]^{1/2} \qquad (5\text{-}6)$$

where ψ_{max} is the maximum quantity of the solution to normalize the ψ_i entities. The maximum ψ_{max} is preferred for the normalization instead of using a relative quantity, e.g. ψ_i^{τ}, which would lead to too strong estimates especially at starting times or if converging to steady state.

5.2.2 Absolute L_1 integral error norm

This norm represents an average of the error in the solution domain:

$$\|e\|_{L_1} = \frac{1}{N\psi_{max}}\sum_{i=1}^{N}|e_i| \qquad (5\text{-}7)$$

It is an alternative to the L_2 RMS norm. Commonly, it should not be the first choice (the RMS norm is often more appropriate).

5.2.3 Maximum L_∞ error norm

This error measure can be useful if focusing on the maximum error occurring in the solution. The maximum norm is defined as

$$\|e\|_{L_\infty} = \frac{1}{\psi_{max}}\max_i |e_i| \qquad (5\text{-}8)$$

It is the strongest measure and should be preferred if the local error is important in the numerical process (*'scheme listens to each sound'*).

5.2.4 Normalization by using the maximum quantity ψ_{max}

The maximum quantity ψ_{max} of the solution $\psi = \psi(x, t)$ is used to normalize the error measures (5-6) to (5-8), where x is the space coordinate and t is the simulation time. Its value is determined internally by FEFLOW at the beginning of each simulation run. In practical simulations ψ_{max} is no more changed during the simulation time t and during the iteration progress τ. Accordingly, ψ_{max} is determined at initial time and for an initial distribution of ψ:

$$\psi_{max} = \max_i \psi_i^0(x, 0) \qquad (5\text{-}9)$$

for $i = 1, ..., N$.

The normalization in form of $\psi_i / \psi_{max}(i = 1, ..., N)$ provides dimensionless measures in (5-6) to (5-8). The input of user-specified error tolerances in FEFLOW is correspondingly dimensionless. In determining the maximum quantity ψ_{max} at initial time via (5-9) the user has to notify the following:

• ψ_{max} is computed at beginning of a simulation run which is based on the initial distribution of ψ and boundary conditions at initial time of the problem, and remains unchanged during the progressing time and iteration. If the problem is not source-free (e.g., the presence of a groundwater recharge, a mass or heat supply) or the boundary conditions are time-dependent, the actual maximum of the solution can increase above the magnitude of ψ_{max} during the time.

• If ψ_{max} becomes a larger value (e.g., a hydraulic head h is referenced to a larger elevation) small derivations in the solutions are scaled down as normalized quantities. For instance a derivation of 1 mm in the hydraulic head leads to different dimensionless error measures if ψ_{max} is different: $(5.001 - 5)/30 = 3.333 \cdot 10^{-5}$, where $\psi_{max} = 30$ m versus $(5.001 - 5)/3000 = 3.333 \cdot 10^{-7}$ when $\psi_{max} = 3000$ m. With other words, if a dimensionless error tolerance amounts for instance to a magnitude of 10^{-4}, the tolerated derivation in the hydraulic head is then 3 mm for the case of $\psi_{max} = 30$ m and 30 cm for the case of $\psi_{max} = 3000$ m

• To avoid a division by zero in the normalization, ψ_{max} is automatically set to 100 if the maximum quantity is detected zero.

5. Error norms used in FEFLOW

About the difference between the convective form and the divergence form of the transport equation

6

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

6.1 Basic Transport Equation

From conservation principles[1] we have the following transport equation for a property ψ

$$\frac{\partial}{\partial t}(R\psi) + \nabla \cdot (q\psi) - \nabla \cdot (D \cdot \nabla \psi) = Q \qquad \text{in} \quad \Omega \quad (6\text{-}1)$$

where

ψ = scalar transport quantity;
R = retardation;
q = velocity field;
D = dispersion tensor;
Q = sink/source;

This formulation is called *divergence form*. The equation (6-1) can be transformed (simplified) by incorporating the continuity equation[1]. It yields

$$R_d \frac{\partial \psi}{\partial t} + q \cdot \nabla \psi - \nabla \cdot (D \cdot \nabla \psi) = Q_\psi \qquad \text{in} \quad \Omega \quad (6\text{-}2)$$

where

$R_d = \partial(R\psi)/\partial\psi$ derivative term of retardation;
$Q_\psi = -\psi Q_\rho + Q$ an additional source term in which Q_ρ is the source term in the continuity equation;

This is called the *convective form*, which is more common and usually applied in practical modeling. The main difference lies in the convective terms. While Eq. (6-1) has a divergence expression $\nabla \cdot (q\psi)$ the transport equation (6-2) involves a more convenient gradient relationship $q \cdot \nabla\psi$ for the convective term. Both transport equations are physically equivalent, but they lead to different formulations of boundary conditions[3] in their discretized forms as shown below.

6.2 Standard Boundary Conditions

Let us denote the boundary of the domain Ω by $\partial\Omega = \Gamma_1 \otimes \Gamma_2$ where Γ_1 and Γ_2 are two disjoint portions of the total boundary, $\partial\Omega$. In the general case following formulations for boundary conditions (BC's) occur, viz.,

$$\left.\begin{array}{ll} \psi = \psi_1 & \text{on} \quad \Gamma_1 \\ -\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi) + a(\psi_2 - \psi) = b & \text{on} \quad \Gamma_2 \end{array}\right\} \quad (6\text{-}3)$$

where on Γ_1 we have Dirichlet BC and on Γ_2 it represents a more general form of a Robin type BC in which more specific Neumann and Cauchy type BC's as used in FEFLOW[1] are involved. If $a = 0$ a Neumann BC of 2nd kind results, while for $b = 0$ a common Cauchy BC of 3rd kind is given. In (6-3) $\boldsymbol{n}$ corresponds to the normal unit vector (positive outward), ψ_1 and ψ_2 are prescribed boundary values of ψ on Γ_1 and Γ_2, respectively.

6.3 Weak Form of the Convective Form (6-2)

The finite element formulation is based on the following weak form of the transport equation (6-2). Introducing the spatial weighting function w we get

$$\int_{\Omega} w\left(R_d \frac{\partial \psi}{\partial t} + \boldsymbol{q} \cdot \nabla \psi\right) = \int_{\Omega} w[\nabla \cdot (\boldsymbol{D} \cdot \nabla \psi) + Q_\psi] \quad (6\text{-}4)$$

We can invoke the following identity (partial integration) applied to the 'dispersive' part

$$\int_{\Omega} \nabla \cdot [w(\boldsymbol{D} \cdot \nabla \psi)] = \int_{\Omega} w[\nabla \cdot (\boldsymbol{D} \cdot \nabla \psi)] \quad (6\text{-}5)$$
$$+ \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla \psi)$$

and rewrite (6-4) as follows:

$$\int_{\Omega} \left[w\left(R_d \frac{\partial \psi}{\partial t} + \boldsymbol{q} \cdot \nabla \psi\right) + \nabla w \cdot (\boldsymbol{D} \cdot \nabla \psi)\right] = \quad (6\text{-}6)$$
$$\int_{\Omega} \{\nabla \cdot [w(\boldsymbol{D} \cdot \nabla \psi)] + wQ_\psi\}$$

Now, applying the divergence theorem (Green's theorem) on the 'dispersive' term on the RHS of (6-6)

$$\int_{\Omega} \nabla \cdot [w(\boldsymbol{D} \cdot \nabla \psi)] = \int_{\partial \Omega} w\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi) \quad (6\text{-}7)$$

and obtain for (6-6)

$$\int_{\Omega} \left[w\left(R_d \frac{\partial \psi}{\partial t} + \boldsymbol{q} \cdot \nabla \psi\right) + \nabla w \cdot (\boldsymbol{D} \cdot \nabla \psi)\right] = \quad (6\text{-}8)$$
$$\int_{\Omega} wQ_\psi + \int_{\partial \Omega} w\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi)$$

We can easily insert the Robin-type BC (6-3) into (6-7) to get

$$\int_{\partial \Omega} w\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi) = \int_{\Gamma_2} w[a(\psi_2 - \psi) - b] \quad (6\text{-}9)$$

By inserting (6-9) into (6-8) the resulting weak form for the finite element solutions is finally given by

$$\int_{\Omega} \left[w\left(R_d \frac{\partial \psi}{\partial t} + \boldsymbol{q} \cdot \nabla \psi\right) + \nabla w \cdot (\boldsymbol{D} \cdot \nabla \psi)\right] + \int_{\Gamma_2} wa\psi = \quad (6\text{-}10)$$
$$\int_{\Omega} wQ_\psi + \int_{\Gamma_2} w(a\psi_2 - b)$$

It should be noticed that with this BC formulation the

normal 'dispersive' flux is expressed. For instance, if $a = b = 0$, known as the natural (Neumann) condition, then the boundary is impervious for the dispersive flux $\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi) = 0$, what does not mean that the boundary is impervious for the normal 'convective' flux $\mathbf{n} \cdot \mathbf{q}\psi$ on the boundary. But we shall see next that this becomes possible in using the divergence form.

6.4 Weak Form of the Divergence Form (6-1)

The weak formulation for the transport equation (6-1) yields

$$\int_\Omega w \left[R_d \frac{\partial \psi}{\partial t} + \nabla \cdot (\mathbf{q}\psi) \right] = \int_\Omega w [\nabla \cdot (\mathbf{D} \cdot \nabla \psi) + Q] \quad (6\text{-}11)$$

Now we see there is one important difference to the above convective form (6-4). Beside the 'dispersive' term we have also to integrate the 'convective' term by parts, viz.,

$$\int_\Omega w \nabla \cdot (\mathbf{q}\psi) = \int_\Omega \nabla \cdot (w\mathbf{q}\psi) - \int_\Omega \psi \mathbf{q} \cdot \nabla w \quad (6\text{-}12)$$

Using the divergence theorem we obtain

$$\int_\Omega \nabla \cdot (w\mathbf{q}\psi) = \int_{\partial\Omega} w\psi \mathbf{n} \cdot \mathbf{q} \quad (6\text{-}13)$$

where a new boundary integral term appears.

Applying this formulation together with the remaining terms (analogous to above) we obtain the following

weak form

$$\int_\Omega \left[w R_d \frac{\partial \psi}{\partial t} - \psi \mathbf{q} \cdot \nabla w + \nabla w \cdot (\mathbf{D} \cdot \nabla \psi) \right] = \quad (6\text{-}14)$$

$$\int_\Omega wQ + \int_{\partial\Omega} w\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi - \mathbf{q}\psi)$$

There is a difference between (6-14) and (6-8) regarding the boundary term for $\partial\Omega$. Would we express the normal convective boundary flux as

$$\int_{\partial\Omega} w\psi \mathbf{n} \cdot \mathbf{q} = \int_{\Gamma_2} w\psi \mathbf{n} \cdot \mathbf{q} \quad (6\text{-}15)$$

and using the Robin BC type (6-3) the weak form of the divergence form (6-14) takes the expression

$$\int_\Omega \left[w R_d \frac{\partial \psi}{\partial t} - \psi \mathbf{q} \cdot \nabla w + \nabla w \cdot (\mathbf{D} \cdot \nabla \psi) \right] + \int_{\Gamma_2} w(a + \mathbf{n} \cdot \mathbf{q})\psi$$
$$= \int_\Omega wQ + \int_{\Gamma_2} w(a\psi_2 - b) \quad (6\text{-}16)$$

The normal convective flux appears now on the LHS of (6-16) and is equivalent to the following BC's

$$\left.\begin{array}{ll} \psi = \psi_1 & \text{on} \quad \Gamma_1 \\ -\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi - \mathbf{q}\psi) + a(\psi_2 - \psi) = b & \text{on} \quad \Gamma_2 \end{array}\right\} \quad (6\text{-}17)$$

rather than (6-3). In applyimg now (6-17) to the divergence form (6-14) the following weak statement results

$$\int_{\Omega} \left[wR_d \frac{\partial \psi}{\partial t} + \nabla w \cdot (D \cdot \nabla \psi - \psi q) \right] + \int_{\Gamma_2} wa\psi$$

$$= \int_{\Omega} wQ + \int_{\Gamma_2} w(a\psi_2 - b)$$

(6-18)

which is different to the convective weak form (6-10). For instance, if $a = b = 0$ the corresponding natural (Neumann) condition forces the total boundary flux ('dispersive' plus 'convective') to be zero. A nonzero value means that both the dispersive and the convective normal flux should be known (or estimable) on Γ_2. As long as $n \cdot q$ is known there it can be handled as a given 'reaction' term in the LHS of (6-16). But normally, if not, for instance along some outflowing boundary portions, this value cannot be predefined *a priori* as a boundary condition and becomes automatically dropped with the consequence that the boundary becomes convectively impervious.

6.5 Advantages *vs.* Disadvantages

The (default) convective formulation (6-10) together with the BC's

$$\left. \begin{array}{ll} \psi = \psi_1 & \text{on } \Gamma_1 \\ -n \cdot (D \cdot \nabla \psi) + a(\psi_2 - \psi) = b & \text{on } \Gamma_2 \end{array} \right\}$$

(6-19)

is easy to handle. All BC's need not specific considerations. Along boundaries the Neumann type $-n \cdot (D \cdot \nabla \psi) = b$ or the Cauchy type $-n \cdot (D \cdot \nabla \psi) = -a(\psi_2 - \psi)$ are appropriate for a wide range of applications. A disadvantage of the convective

form can arise if the prescription of the 'dispersive' boundary flux is insufficient, e.g., for an intruding contaminant source on a boundary portion with a given rate.

The (optional) divergence form (6-16) is able to conserve the total ('convective' plus 'dispersive') boundary flux on boundaries:

$$\left. \begin{array}{ll} \psi = \psi_1 & \text{on } \Gamma_1 \\ -n \cdot (D \cdot \nabla \psi - q\psi) + a(\psi_2 - \psi) = b & \text{on } \Gamma_2 \end{array} \right\}$$

(6-20)

Here, the Neumann type condition is $-n \cdot (D \cdot \nabla \psi - q\psi) = b$, while a Cauchy type BC takes the form $-n \cdot (D \cdot \nabla \psi - q\psi) = -a(\psi_2 - \psi)$. This is sometimes advantageous because the total flux BC is satisfied[2] instead of satisfying only the 'dispersive' flux. However, on outflowing boundaries $n \cdot q$ is often unknown and such a BC type requires a specific handling as described in the following section.

As the sum, the convective form is much easier to handle and sufficient for the most applications, while the divergence form represents a totally conserved formulation, however, it needs specific techniques and additional effort for boundary conditions at outflowing portions.

6.6 Handling of Outflowing Boundaries for the Divergence Form

At an outflowing boundary the Robin-type boundary condition

$$-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla\psi - \boldsymbol{q}\psi) + a(\psi_2 - \psi) = b \qquad \text{on} \qquad \Gamma_2 \quad (6\text{-}21)$$

cannot be specified *a priori* because $\boldsymbol{n} \cdot \boldsymbol{q}\psi|_{\Gamma_2}$ is unknown. However, this part can be computed from the flow equation via a postprocessing balance analysis in the following form:

The flow equation

$$S_o \frac{\partial h}{\partial t} - \nabla \cdot (\boldsymbol{K} \cdot \nabla h) = Q_\rho \qquad \text{in} \qquad \Omega \qquad (6\text{-}22)$$

where h is the hydraulic head, S_o is the storage coefficient and $\boldsymbol{K}$ corresponds to the tensor of hydraulic conductivity, leads to the following weak formulation

$$\int_\Omega wS_o \frac{\partial h}{\partial t} + \int_\Omega \nabla w \cdot (\boldsymbol{K} \cdot \nabla h) = -\int_{\partial\Omega} w\boldsymbol{n} \cdot \boldsymbol{q} + \int_\Omega wQ_\rho \quad (6\text{-}23)$$

With the given solution of h Eq. (6-23) can be explicitly evaluated at Γ_2 according to

$$\int_{\Gamma_2} w\boldsymbol{n} \cdot \boldsymbol{q} = \int_\Omega wQ_\rho - \int_\Omega \nabla w \cdot (\boldsymbol{K} \cdot \nabla h) - \int_\Omega wS_o \frac{\partial h}{\partial t} \quad (6\text{-}24)$$

via a *consistent boundary quantity method*[4] to give the unknown boundary flux $\boldsymbol{n} \cdot \boldsymbol{q}|_{\Gamma_2}$. Here, the boundary quantity $\boldsymbol{n} \cdot \boldsymbol{q}$ is obtained from (6-24) by summing up all nodal contributions at the corresponding boundary portion over all adjacent finite elements e

$$\int_{\Gamma_2} \boldsymbol{n} \cdot \boldsymbol{q} = \sum_e \int_{\Gamma_2^e} w\boldsymbol{n} \cdot \boldsymbol{q} \qquad (6\text{-}25)$$

As soon $\boldsymbol{n} \cdot \boldsymbol{q}$ becomes known at Γ_2 the Robin-type boundary condition (6-21) can be replaced by such a type (6-19)

$$-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla\psi) + a(\psi_2 - \psi) = b \qquad \text{on} \qquad \Gamma_2 \quad (6\text{-}26)$$

which is also used for the convective form of the transport equation. With other words, at outflowing boundaries the divergence form turns on a 'diffusive' Neumann-type boundary condition (6-26), e.g., $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla\psi) = b$ instead of $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla\psi - \boldsymbol{q}\psi) = b$.

Finally, at outflowing boundaries the weak form of the transport equation (6-16) is now used as

$$\int_\Omega \left[wR_d \frac{\partial\psi}{\partial t} - \psi\boldsymbol{q} \cdot \nabla w + \nabla w \cdot (\boldsymbol{D} \cdot \nabla\psi) \right]$$

$$+ \int_{\Gamma_2} w \left(a + \underbrace{\boldsymbol{n} \cdot \boldsymbol{q}}_{\text{given from Eq. (24)}} \right) \psi \qquad (6\text{-}27)$$

$$= \int_\Omega wQ + \int_{\Gamma_2} w \left(a\psi_2 - \underbrace{b}_{\text{according to Eq. (25)}} \right)$$

with $\boldsymbol{n} \cdot \boldsymbol{q}$ as a given 'reaction term' computed from (6-24) and (6-25), and b as a 'diffusive-type' boundary condition based on the formulation of Eq. (6-26). In this way, the divergence form (6-14) becomes applicable for all practically important boundary conditions: While at inflowing boundaries a total mass flux condition can be preferable, at outflowing boundaries only the diffusive/dispersive outgoing needs to be specified, where $\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla\psi) = 0$ appears as a *natural boundary condition* for outflowing situations.

6.7 An Example

Considering the 2D vertical flow domain as shown in Fig. 6.1 describing the contaminant influence on a discharging aquifer from a leaked deposit, the differences between the convective and the divergence form of the transport equations become apparent. The cross-sectional domain has a length of 1000 m and a height of 40 m. The aquifer is discharged from left to right. The deposit contacts the aquifer on top over a length of 50 m, where contaminant matter releases. All parameters and relationships are displayed in Fig. 6.1.

The boundary conditions for the flow problem are summarized in Tab. 6.1. Accordingly, the discharge amounts to a total flow balance of $0.1 \cdot 40 + 0.0125 \cdot 50 = 4.625 \text{ m}^3/\text{d/m}$, which releases through the boundary D-E.

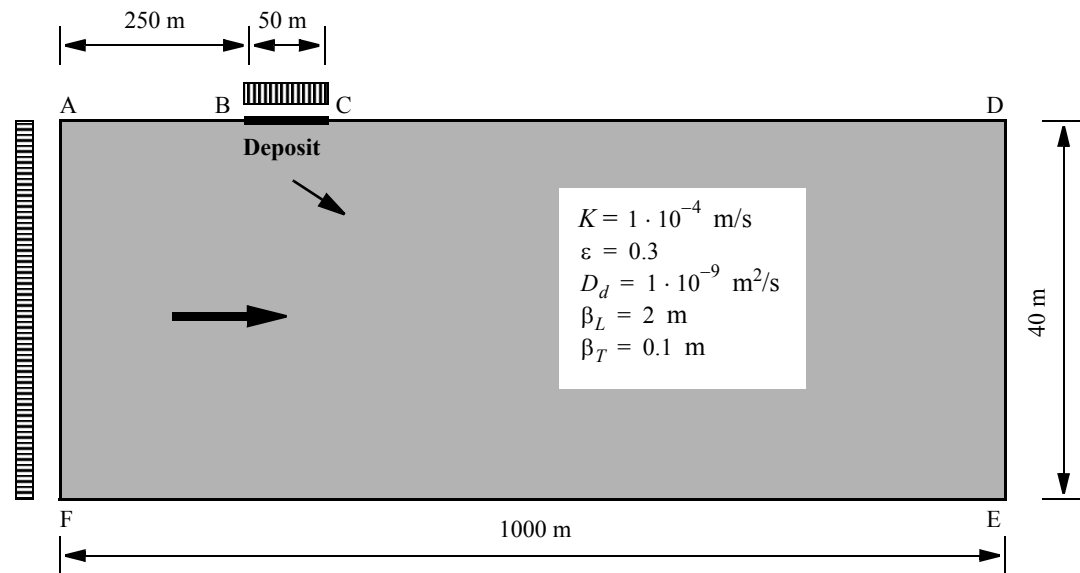**Figure 6.1** Cross-sectional domain of study (K - isotropic hydraulic conductivity, ε - porosity, D_d - molecular diffusion, β_L - longitudinal dispersivity, β_T - transverse dispersivity).

Table 6.1 Flow boundary conditions

Section	Type	Value	Unit	Comment
A-B	-	-	-	unspecified (impervious)
B-C	Neumann	$\mathbf{n} \cdot \mathbf{q}$ = -0.0125	m/d	influx
C-D	-	-	-	unspecified (impervious)
D-E	Dirichlet	$h = 40$	m	pervious boundary (outflux)
E-F	-	-	-	unspecified (impervious)
A-F	Neumann	$\mathbf{n} \cdot \mathbf{q}$ = -0.1	m/d	influx

The boundary conditions for the contaminant mass transport differ between the convective and the divergence forms as summarized in Tabs. 6.2 and 6.3. The contaminant release from the deposit is modeled by a given rate. This rate represents a dispersive influx condition $\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi)$ for the convective form according to Eq. (6-19) and a total (dispersive plus convective) rate $\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi - \mathbf{q} \psi)$ for the divergence form according to Eq. (6-20).

Table 6.2 Boundary conditions for the convective form according to Eq. (6-19)

Section	Type	Value	Unit	Comment
A-B	-	-	-	unspecified (impervious for dispersive fluxes $-\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi) = 0$)
B-C	Neumann $(a = 0)$	$-\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi)$ = -0.0125	g/m^2/d	predefined (dispersive) influx of contaminant
C-D	-	-	-	unspecified (impervious for dispersive fluxes $-\mathbf{n} \cdot (\mathbf{D} \cdot \nabla \psi) = 0$)

Table 6.2 Boundary conditions for the convective form according to Eq. (6-19) (continued)

Section	Type	Value	Unit	Comment
D-E	-	-	-	unspecified (impervious for dispersive fluxes $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi) = 0$, however, the boundary is convectively pervious, compare Tab. 6.1)
E-F	-	-	-	unspecified (impervious for dispersive fluxes $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi) = 0$)
A-F	Dirichlet	$\psi = \psi_1 = 0$	g/m^3	predefined concentration (entering freshwater)

Table 6.3 Boundary conditions for the divergence form according to Eq. (6-20)

Section	Type	Value	Unit	Comment
A-B	-	-	-	unspecified (impervious for total fluxes $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi - \boldsymbol{q}\psi) = 0$)
B-C	Neumann ($a = 0$)	$-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi - \boldsymbol{q}\psi) = -0.0125$	g/m^2/d	predefined (total) influx of contaminant
C-D	-	-	-	unspecified (impervious for total fluxes $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi - \boldsymbol{q}\psi) = 0$)
D-E	-	-	-	unspecified, however, specific handling as outflowing boundary in setting $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi) = 0$ as an impervious condition for dispersive fluxes

Table 6.3 Boundary conditions for the divergence form according to Eq. (6-20) (continued)

Section	Type	Value	Unit	Comment
E-F	-	-	-	unspecified (impervious for total fluxes $-\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi - \boldsymbol{q} \psi) = 0$)
A-F	Dirichlet	$\psi = \psi_1 = 0$	g/m^3	predefined concentration (entering freshwater)

The computed stationary contaminant plume for the convective form is exhibited in Fig. 6.2. The counterpart of the divergence form is shown in Fig. 6.3. Qualitatively, both plumes have similar characteristics, but in their quantity the patterns are fairly depart of each other.

In the convective form the Neumann condition for the contaminant release at the deposit is mimicked by a dispersive flux $\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla \psi)$ with a magnitude of 0.0125 $g/m^2/d$. Notice, for this formulation the convective influx at this boundary portion is not defined. To realize this given rate via a dispersive mechanism (i.e., driven by a concentration gradient $\nabla \psi$ and controlled by the dispersion $\boldsymbol{D}$) the concentration gradient results automatically, where a certain concentration magnitude appears at the deposit border section. In the present case, the concentration increases to a maximum of about 7.17 g/m^3. In a budget analysis for the convective form a total amount of 3.45 g/m/d contaminant mass flux appears. What is the reason for this enlarged contaminant mass flux? In the convective form only the dispersive part is defined at the boundary while the convective part remains undefined. Expectedly, the dispersive magnitude is exactly 0.625 g/m/d. With the given dispersivity parameters (Fig. 6.1) concentration profiles along the deposit boundary result as shown in Fig. 6.4. This concentrations multiplied with the flow rate of 0.0125 m/d (Tab. 6.1) must be additionally convected through the boundary. Because the concentrations are relatively high the convective part must be large. Notice, it should become clear that this convective form must fail for the given boundary conditions if the dispersivity (and diffusion) becomes smaller and smaller. Then, the concentration gradient tends to an infinitely large value with the result that the convective part of the boundary flux becomes infinitely large.

Unlike this, the divergence form echoes exactly the distribution and balance quantities which should be practically expected. The budget analysis results a total contaminant release of 0.625 g/m/d through the deposit boundary. Consequently, the concentration magnitudes are significantly smaller (maximum value is only about 1 g/m^3 at the deposit boundary) as shown in Fig. 6.4. This means the net contaminant mass release for the divergence form is about six times smaller than for the convective form, which is also indicated by the departing concentration profiles of this order (Fig. 6.4).

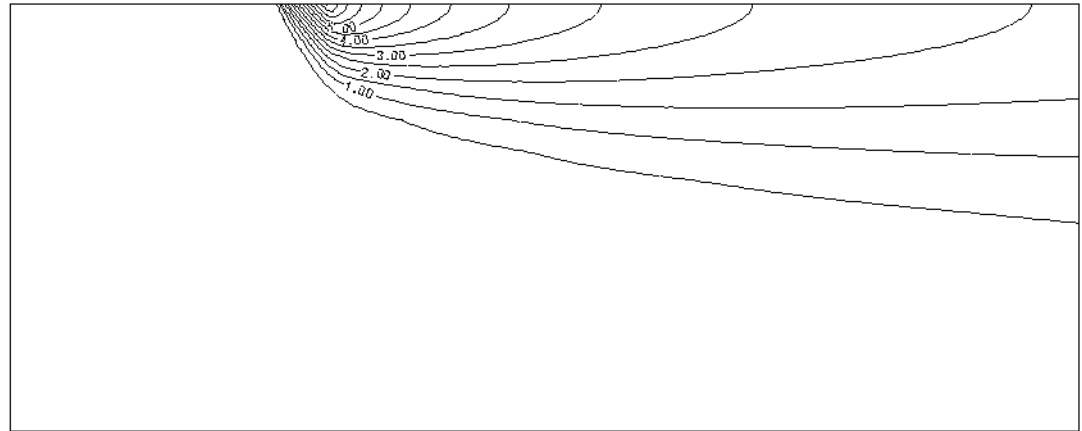# 6. About the difference between the convective form and the divergence form of the transport equation

Figure 6.2 Computed stationary contaminant distribution for the convective form (exaggeration 10 : 1).
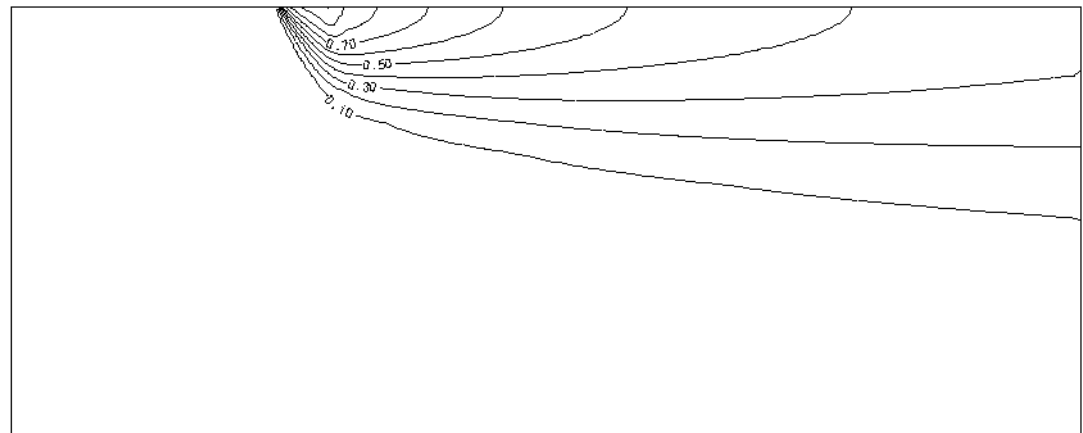

Figure 6.3 Computed stationary contaminant distribution for the divergence form (exaggeration 10 : 1).

The differences between the convective and the divergence forms result from the different meaning of boundary conditions for fluxes (Neumann-type or Cauchy-type boundary conditions) of contaminants. It should be obvious that no differences appear if the contaminant source is modeled by a Dirichlet-type boundary condition where the concentration is fixed. It is to be noted that the same relationships have to be considered for heat transport phenomena.
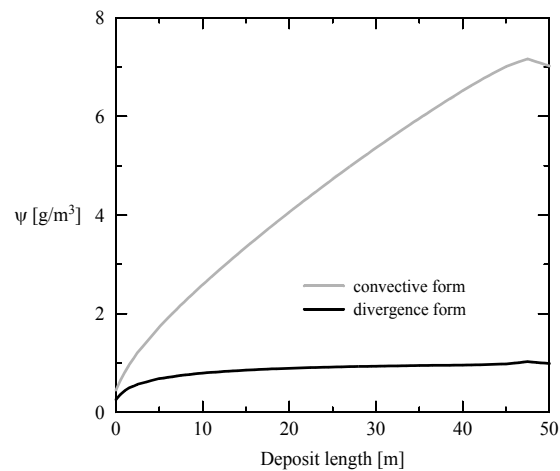
Figure 6.4 Concentration profiles computed along the deposit boundary in dependence on the convective and divergence form.

3. Galeati, G. & Gambolatti, G., On the boundary conditions and point sources in the finite element integration of the transport equation, *Water Resources Research* **25** (1989) 5, 847-856.
4. Gresho, P. M., Lee, P. L. & Sani, R. L., The consistent method for computing derived boundary quantities when the Galerkin FEM is used to solve thermal and/or fluids problems, *Numerical Methods in Thermal Problems*, Vol. 2, Pineridge Press, 1981, 663-675.

References

1. Diersch, H.-J. G., *FEFLOW - Physical basis of modeling.* Reference Manual - Part I, WASY Ltd., Berlin, 2002.
2. Pinder, G. & Gray, W., *Finite element simulation in surface and subsurface hydrology.* Academic Press, 1977.

6. About the difference between the convective form and the divergence form of the transport equation

About the formulation of hydraulic head boundary (potential) conditions for fluid density-dependent groundwater problems

7

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

7.1 Problem Description

In formulating boundary conditions (B.C.) for the groundwater flow the prescription of a hydraulic head (1st kind Dirichlet-type) B.C. at a given boundary portion is a common task. However, in modeling density-dependent problems such as saltwater intrusion or geothermal processes these 1st kind hydraulic head B.C.'s have to consider the specific definition of the hydraulic head (potential)[1]. A typical example is the saltwater intrusion from a sea into a coastal aquifer as schematized in Fig. 7.1. While on the land side of the aquifer a freshwater discharge can be prescribed (e.g., by a 2nd kind Neumann flux-type B.C.), at the sea side the boundary is formed by a given hydraulic head distribution. This hydraulic potential is measured in form of the piezometric head at the sea which is related to the actual fluid density of saltwater ρ_s.
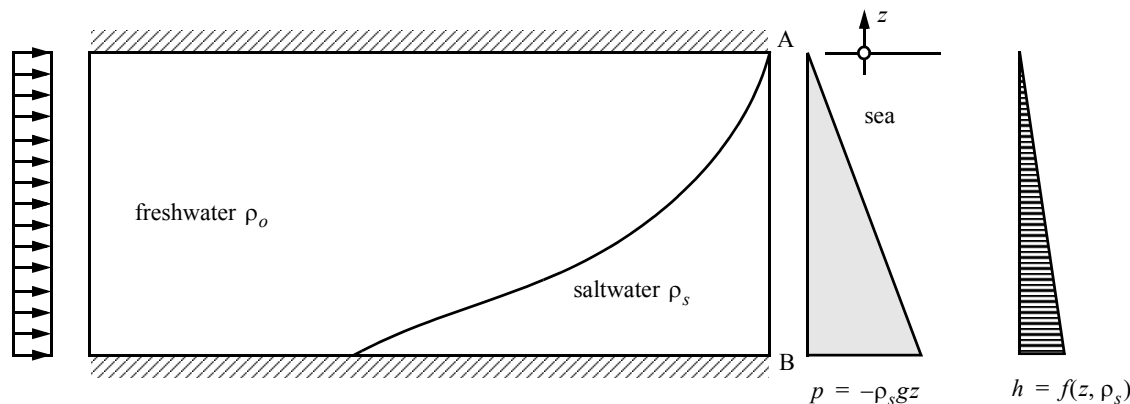

Figure 7.1 Saltwater intrusion in a coastal aquifer with related boundary conditions.

7.2 The Reference Hydraulic Potential

For the groundwater flow equations[1] FEFLOW prefers the *hydraulic head* h, instead of the pressure p, as the primary variable. As a consequence, the corresponding B.C.'s of 1st kind (Dirichlet-type)

$$h = h^R \qquad \text{at} \qquad \Gamma_{\widehat{AB}} \qquad (7\text{-}1)$$

have to be expressed according to the definition of the hydraulic potential h. As described in[1] the hydraulic head variable h must be appropriately related to a *reference fluid density* ρ_o, viz.,

$$h = \frac{p}{\rho_o g} + z \qquad (7\text{-}2)$$

Commonly, for saltwater intrusion problems the reference density ρ_o refers to the *freshwater*.

7.3 Reference Potential from Measured Heads

A measurement of a piezometric head is normally related to the actual density of the groundwater. It can be expressed by

$$h_s = \frac{p}{\rho_s g} + z \qquad (7\text{-}3)$$

where ρ_s is the density of groundwater at a known salinity C: $\rho_s = \rho_s(C)$. It should be clear that the head h_s cannot be directly used as a boundary condition. Instead, it has to be transformed to the reference hydraulic head h, Eq. (7-2). This can be simply done under considering the following relationships:

Expanding Eq. (7-3) by ρ_o

$$h_s = \frac{p}{\rho_o g} \frac{\rho_o}{\rho_s} + z \qquad (7\text{-}4)$$

we get if introducing Eq. (7-2)

$$h_s = \frac{\rho_o}{\rho_s} h + \left(1 - \frac{\rho_o}{\rho_s}\right) z \qquad (7\text{-}5)$$

and finally

$$h = \frac{\rho_s}{\rho_o} h_s - \left(\frac{\rho_s - \rho_o}{\rho_o}\right) z \qquad (7\text{-}6)$$

Now, introducing the *density difference ratio* α as[1]

$$\alpha = \frac{\rho_s - \rho_o}{\rho_o} \qquad (7\text{-}7)$$

Eq. (7-6) can also be written in the form

$$h = (1 + \alpha) h_s - \alpha z \qquad (7\text{-}8)$$

Equations (7-6) or (7-8) have to be used to calculate the hydraulic head h from piezometric heads h_s which have been measured at a known saltwater density ρ_s (at known salinity C).

7.4 Hydrostatic Condition

Let us consider the pressure distribution in the vertical z-direction of gravity g under hydrostatic conditions. We assume that the density $\rho = \rho(z)$ is varying linearly in the depth as shown in Fig 7.2:

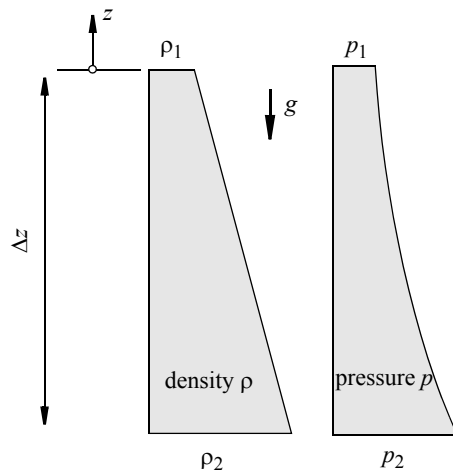$$\rho = \rho_1 + \frac{(\rho_1 - \rho_2)}{\Delta z} z \qquad (7\text{-}9)$$

Figure 7.2 Hydrostatic condition in a depth of Δz under a linear density gradient.

The fluid is hydrostatic for the vertical problem if

$$\left. \begin{array}{c} \dfrac{dp}{dz} = -\rho g \\[2em] p = p(z) = -g \displaystyle\int_{z_1}^{z} \rho(\theta) d\theta \end{array} \right\} \qquad (7\text{-}10)$$

which yields with (7-9)

$$p = p_1 - g\left(\rho_1 z + \frac{\rho_1 - \rho_2}{2\Delta z} z^2\right) \qquad (7\text{-}11)$$

The hydraulic head h (7-2) related to the reference density ρ_o is then

$$h = h_1 - \left(\frac{\rho_1 - \rho_0}{\rho_0}\right) z - \frac{1}{2\Delta z}\left(\frac{\rho_1 - \rho_2}{\rho_0}\right) z^2 \qquad (7\text{-}12)$$

At boundaries where hydrostatic conditions can be imposed two cases are commonly of interest: (1) a constant saltwater density in the depth and (2) a linear increase of density as typical in a transition zone. Both cases are illustrated in Fig. 7.3.
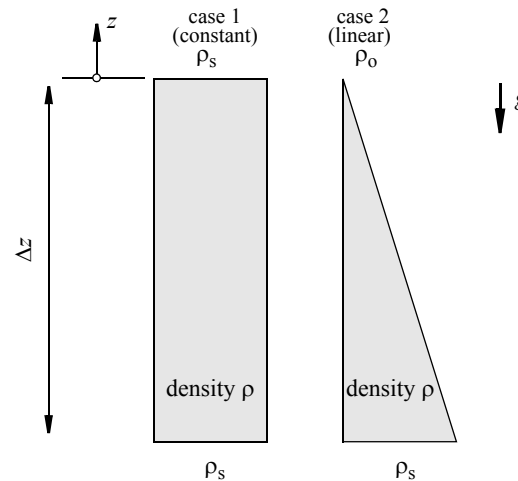

Figure 7.3 Two interesting density profiles for a hydrostatic boundary condition.

From (7-12) we obtain with $\rho_1 = \rho_2 = \rho_s$ for the case of a constant saltwater density (case 1)

$$h = h_s - \alpha z \tag{7-13}$$

For the case 2 with $\rho_1 = \rho_0$, $\rho_2 = \rho_s$ we get from (7-12) for a linear saltwater density (case 2)

$$h = h_o + \frac{\alpha}{2\Delta z} z^2 \tag{7-14}$$

The hydraulic head at the depth $z = -\Delta z$ is then $h = h_s + \alpha\Delta z$ for the constant density and $h = h_o + \frac{\alpha}{2}\Delta z$ for the linear density relationship.

7.5 Examples

7.5.1 Boundary with constant density

Referring to the saltwater intrusion problem of Fig. 7.1 we assume that water table of the sea is given by $h_s = 0$, that means we choose $z = 0$ at the free surface of the sea where we have $p = 0$. We aim at finding a corresponding hydraulic head condition for h which should be imposed on the sea side of the model domain $\widehat{AB}$ (cf. Fig. 7.1). Let us assume we have a density of the seawater of $\rho_s = 1.029$ kg/l. The freshwater has $\rho_o = 1$ kg/l, accordingly one gets $\alpha = 0.029$. The thickness of the aquifer should be 20 m ($0 \leq z \leq -20$m).

Applying Eq. (7-8) or (7-13) we find along the vertical boundary portion $\widehat{AB}$ the following distribution for the hydraulic head h:

$$h\big|_{\widehat{AB}} = -\alpha z = -0.029z \tag{7-15}$$

Accordingly, at the point A (top of the aquifer, Fig. 7.1) we have to set $h = 0$ and at the point B (bottom of the aquifer) we must set $h = 0.58$ m. Between these nodes the distribution of h is linear so as seen by Eq. (7-15).

REMARK 1: The assignment of the hydraulic head h has to be consistent with the definition of the z-coordinate in the direction of gravity according to Eq. (7-8). In contrast to the above example if we would define $z = 0$ at the aquifer bottom (at point B in Fig. 7.1), h_s has to be chosen as 20 m and Eq. (7-8) results $h\big|_{\widehat{AB}} = 20.58 - 0.029z$ [m], which represents an equivalent prescription of the hydraulic head B.C.'s.

REMARK 2: Similar considerations are required for thermal (thermohaline) problems if a hydraulic head condition is to be prescribed at a boundary where the temperature T (and salinity C) are input. In this case the density ρ_s is deemed to be the value for the measured T (and C): $\rho_s = \rho_s(T, C)$. For instance, using a linear thermal expansion of the density[1] we have for a thermohaline situation (both T and C are given here) the following relationship

$$\frac{\rho_s - \rho_o}{\rho_o} = \alpha - \beta(T_s - T_o) \tag{7-16}$$

and Eq. (7-6) yields

$$h = [1 + \alpha - \beta(T_s - T_o)]h_s - [\alpha - \beta(T_s - T_o)]z \tag{7-17}$$

in which T_s is the temperature to be prescribed at the

boundary $\Gamma_{\widehat{AB}}$, T_o is the reference temperature related to the reference density ρ_o, and β is a known linear thermal expansion coefficient. (Notice, for purely thermal problems $\alpha = 0$ in Eq. (7-17)). To demonstrate this relationship let us expand the above saltwater intrusion example (Fig. 7.1). We assume the sea water has a temperature of $20°C$ while the freshwater is $6°C$ cold. The β-coefficient is $10^{-4} K^{-1}$, α is again 0.029. Accordingly, the input hydraulic head h at the boundary portion $\widehat{AB}$ varies as

$$h\big|_{\widehat{AB}} = -[\alpha - \beta(T_s - T_o)]z = -0.0276z \qquad (7\text{-}18)$$

7.5.2 Boundary with variable density

Sometimes it is necessary to impose a transition zone at a boundary for a saltwater intrusion process. Such an example is shown in Fig. 7.4, where the density varies linearly through the transition zone with a thickness of Δz_1.
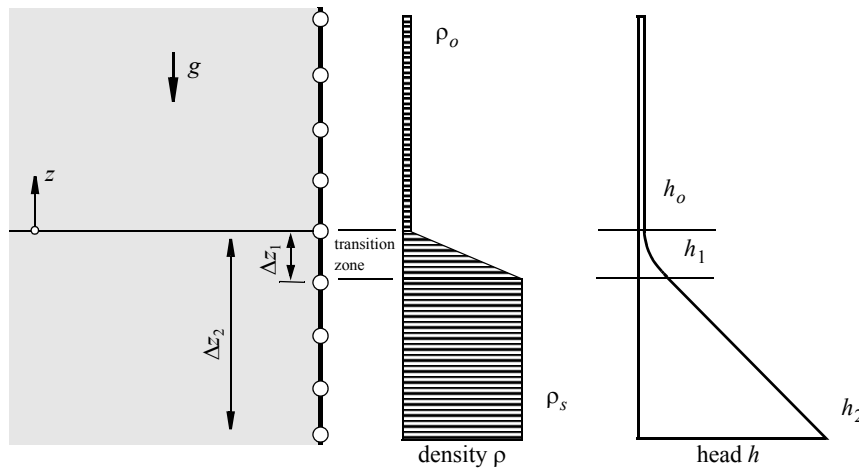
Figure 7.4 Boundary with a predefined saltwater-freshwater transition zone.

At such a boundary a hydraulic head condition $h = h(z)$ has to be imposed. From Eqs. (7-14) and (7-13) we obtain the following sample values for the head profile as indicated in Fig. 7.4:

$$\left.\begin{array}{l} h_1 = h_o + \dfrac{\alpha}{2}\Delta z_1 \\[2ex] h_2 = h_1 + \alpha(\Delta z_2 - \Delta z_1) = h_o + \alpha\left(\Delta z_2 - \dfrac{\Delta z_1}{2}\right) \end{array}\right\} \qquad (7\text{-}19)$$

where h_o represents the hydraulic head at the boundary which is related to the freshwater density ρ_o.

Notation

C	=	salinity, saltwater concentration, (ML^{-3});
g	=	gravitational acceleration, (LT^{-2});
h	=	hydraulic head referenced to ρ_o, (L);
h_s	=	measured saltwater piezometric head related to ρ_s, (L);
h_0, h_1, h_2	=	heads at locations, (L);
p	=	pressure, $(ML^{-1}T^{-2})$;
T	=	temperature, (Θ);
T_s, T_o	=	boundary temperature and reference temperature, respectively, (Θ);
z	=	elevation; Cartesian coordinate along acting of gravity, (L);
α	=	density difference ratio, (1);
β	=	linear thermal expansion coefficient, (Θ^{-1});
Δz	=	depth, (L);
Γ	=	boundary portion;
ρ_o	=	reference fluid density; density of freshwater, (ML^{-3});
ρ_s	=	density of saltwater, (ML^{-3});

References

1. Diersch, H.-J. G., *FEFLOW - Physical basis of modeling.* Reference Manual - Part I, WASY Ltd., Berlin, 2002.

An efficient method for computing groundwater residence times

8

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

8.1   Introduction

The computation of residence (travel) times of dissolved solutes in the groundwater body or the determination of the groundwater age is traditionally performed by *particle tracking* methods[1,2,6] based on the advective pore velocity distributions. Unfortunately, such type of technique provides only point-related information about the groundwater age in form of isochrones and, furthermore, neglects effects of hydrodynamic dispersion. A certain expedient can provide *random work* techniques[6]. However, they are again point-related approaches and are often time-consuming, especially for 3D applications, since a large number of particles are needed to obtain representative results for practical requirements.

Recently Goode[5] has proposed an interesting alternative in contrast to the above traditional approaches. Its method is capable of computing the *groundwater age* in a direct manner (practically in one step) at any points of the model domain. Additionally, it includes effects of the advection, diffusion and dispersion processes. Goode's method can be immediately performed by FEFLOW[3] provided the basic variables, parameters and boundary conditions are appropriately chosen. Goode's direct simulation strategy has taken over by Perrochet[7,4] and was successfully applied to practical tasks by using FEFLOW. In the following the *direct age simulation* strategy will be described in some detail, which can be adapted by each FEFLOW user. The main advantages of this approach can be summarized as follows:

- It can be performed by FEFLOW in utilizing the embodied modeling features.
- Both 2D and 3D problems are easily applicable.
- The method is effective and can be simply handled.
- It also includes effects of mechanical dispersion and diffusion.
- It can also be applied to transient flow conditions.

8.2   Transport Equation of the Groundwater Age

For a steady-state flow field the mean residence time ('age') A [d] can be determined from the concen-

tration C [mg/l] of a tracer injected as an impulse at time zero[5]. Accordingly, at any point in the model domain the age A of the groundwater is considered as a concentration-weighted travel time, *viz.*,

$$A = \frac{\int_0^\infty tC dt}{\int_0^\infty C dt} \tag{8-1}$$

where t is the time and C corresponds to the concentration of the tracer. The dissolved concentration of the tracer have to satisfy the law of mass conservation written in form of the advection-dispersion transport equation[5]:

$$\varepsilon\frac{\partial C}{\partial t} + q \cdot \nabla C = \nabla \cdot (D \cdot \nabla C) \tag{8-2}$$

where ε is the porosity, q is the Darcy velocity vector and D represents the tensor of the hydrodynamic dispersion which includes effects of molecular diffusion D_d, longitudinal and transverse dispersivities β_L, β_T, respectively. Multiplying Eq. (8-2) by time, integrating through all times, applying partial integration and inserting the definition (8-1), one finally obtains a transport equation of the following type

$$q \cdot \nabla A - \nabla \cdot (D \cdot \nabla A) = \varepsilon \tag{8-3}$$

Equation (8-3) represents a steady-state transport equation in which the mean age A is the primary variable and the porosity ε appears as an 'age source' term of unit strength on the right-hand side. (It should be noted for a depth-integrated horizontal 2D transport equation the age source is $B\varepsilon$, where B corresponds to the thickness of the aquifer.)

To solve Eq. (8-3) for the age A under a given steady groundwater flow field q appropriate boundary conditions of 1st kind (Dirichlet-type) and 2nd kind (Neumann-type) along inflowing and outflowing boundary sections have to be prescribed in the following manner: At inflowing boundaries the groundwater age A can be usually imposed as a 1st kind boundary condition, for instance, if setting $A = 0$ the age (as a relative time) is considered as the beginning time on such a boundary section. On the other hand, along outflowing boundaries a natural 2nd kind Neumann condition can often be specified as $n \cdot (D \cdot \nabla A) = 0$, i.e., the age in normal direction n to the boundary does not change anymore (for instance typically if groundwater leaves the aquifer and enters surface water).

8.3 Working Steps in FEFLOW

The solution of the age transport equation (8-3) in 2D and 3D can be simply performed by FEFLOW. One utilizes the implemented transport equations which are basically available in terms of either the contaminant mass C or the temperature T variables. Instead, the solution of the mass (or heat) transport equation is mimicked for the age A. The following working steps are now useful:

(1) Specification of a steady-state flow problem in a common manner.

(2) Extension of the problem class to a transport problem. We recommend a mass transport problem under steady-state conditions (i.e., steady flow - steady mass

transport).

(3) Formulation of boundary conditions for the age A, e.g., $A = 0$ at inflowing boundaries and $n \cdot (D \cdot \nabla A) = 0$ at outflowing boundaries.

(4) Specification of the material conditions, where properly the source term of zero order ('sink/source') is to be set as the age source in form of ε (or $B\varepsilon$). (Note, FEFLOW's problem editor provides copy functions which benefits the assignment of the age source from porosity data).

(5) Solution of the steady flow and the age transport equations in one step. The evaluation of the results can be done by the standard tools available in FEFLOW for the concentration, e.g., isoline plotting, 3D visualization, data exporting etc.

There is also a trick[7] if the age computation is required parallel to another transport equation so as needed for density-driven processes. In this case the problem is classified as a thermohaline problem in which flow, mass and heat transport are simultaneously simulated. Either the mass or the heat transport equation can then be used as the age transport equation. This allows an new approach to the analysis of residence times for complex flow situations which were impossible to date.

8.4 Demonstrative Example

Let us consider two aquifers which are separated by an aquitard. The ratio of the hydraulic conductivities of the aquitard to the both aquifers is amounted to 1 :

1000. Furthermore, we assume a disturbance in the aquitard in form of local 'hydrogeologic window', through which the lower aquifer can be threatened. The hydrogeologic window should have the same conductivity than the aquifer. On top of the upper aquifer groundwater recharge is input. The following questions arise: At which travel times do surface-entering contaminants arrive the groundwater at different depths? How is the influence of mechanical dispersion and diffusion? What are the differences between a 2D and 3D modeling of the hydrogeologic window in the aquitard?

We start with a 2D modeling. Figure 8.1 displays the cross-section of a 2D vertical model, where the depicted steady-state pathlines and isochrones are obtained by the traditional particle tracking approach available in FEFLOW too.
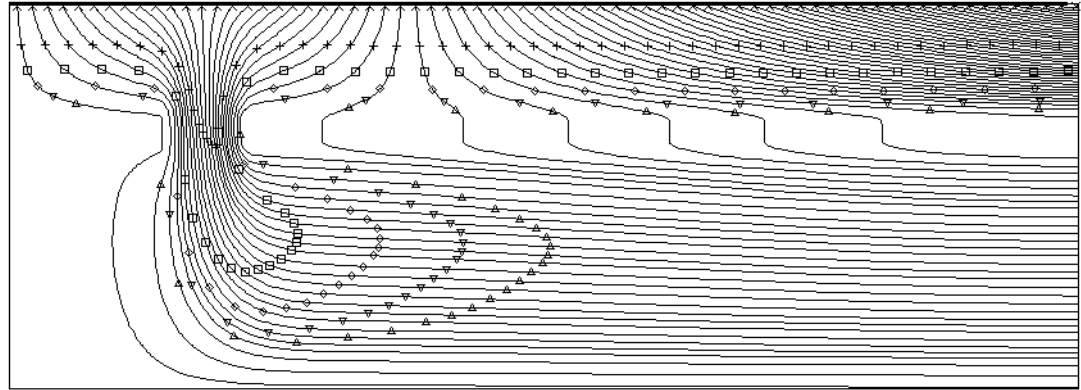
Figure 8.1 2D (exaggerated) cross-sectional domain with pathlines and isochrones at 10, 20, 30, 40 and 50 years simulated by FEFLOW's traditional particle tracking approach.

Now, the computation of the same problem via the direct simulation of the groundwater age results a distribution of ages A for 10, 20,. 30, 40 and 50 years in an isoline plot as shown in Fig. 8.2. Along the upper boundary on which groundwater recharge is entering, the age A is set to zero. A comparison of the results (Fig. 8.2) with the particle tracking analysis (Fig. 8.1) reveals a close agreement. Both simulations are performed on the same mesh. The comparison between the traditional particle tracking and Goode's direct age simulation requires negligible dispersion. Accordingly, the results of Fig. 8.2 have been achieved by suppressing the dispersivities ($\beta_L = \beta_T = 0$). To stabilize the solution for the direct age simulation a streamline-upwind method was used.

The effect of the hydrodynamic dispersion can be seen in Fig. 8.3 for the depicted age distribution. Compared to the case without dispersion (Fig. 8.2) two main differences appears which can be of certain importance for practical applications: (1) If dispersion is considered the age is reduced in locations which are mainly advectively affected as can be seen in the flow region directly below the hydrogeologic window. It means a longer travel time is required before a recharge influence starting from the surface travels to a point in the lower aquifer. This is caused by the dispersion, where flow particles have to go a longer pathway within the void space. However, a contrary effect can be observed at the aquifer-aquitard contact zone. (2) The age increases at the occurrence of hydrodynamic dispersion (including diffusion) in impermeable or low-permeable parts. This reveals physicochemical effects on the travel times of the groundwater below the aquitard which cannot be studied by common particle tracking methods.
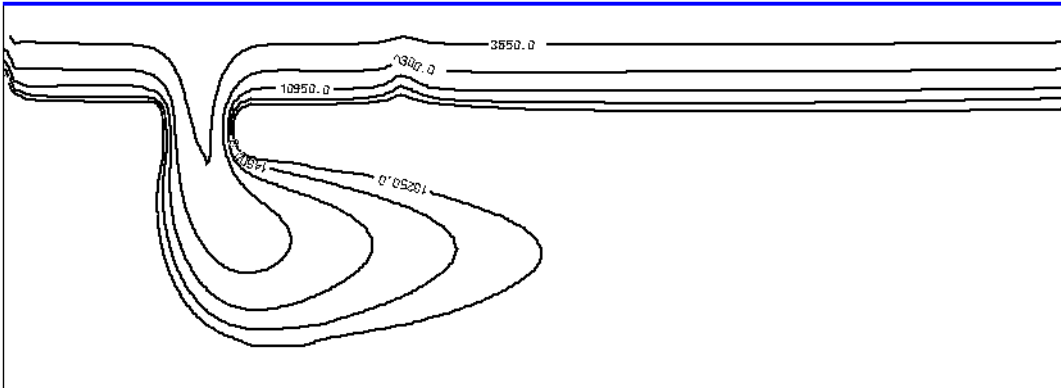
Figure 8.2 Computed ages for 10, 20, 30, 40 and 50 years by the direct simulation of groundwater age (exaggerated cross-section), without mechanical dispersion $\beta_L = \beta_T = 0$.
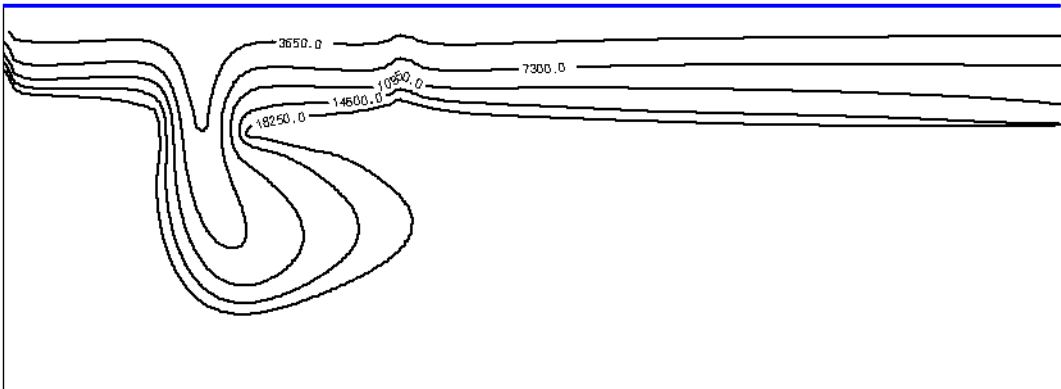

Figure 8.3 Computed ages for 10, 20, 30, 40 and 50 years by the direct simulation of groundwater age (exaggerated cross-section), with mechanical dispersion.

The use of Goode's direct age simulation is simply possible in the same way for 3D application, which will be shown next along the above two-aquifer-aquitard problem. Now we assume that the hydrogeologic window in the aquifer has a three-dimensional extent. If we use the traditional particle tracking method also available for 3D in FEFLOW we can find pathlines and isochrone patterns so as displayed in Fig. 8.4. Since each particle tracking event is always related to single starting point one needs many points, especially in 3D, to get a possibly closed representation and to record (hopefully) all critical locations. For complex applications this leads immediately to a 'chaos' of lines and markers in the 3D space. In contrast to that, the proposed direct age simulation does not suffer in such difficulties. Here, the groundwater age represents a scalar quantity computed at each node of a mesh. It can be evaluated by using the available postprocessing tools, for instance, isolines or fringes in slices, through arbitrary cross-sections and 3D displays so as exemplified in Fig. 8.5 showing the age distribution for 50 years in form of a 3D isosurface.

times in 3D applications. It should be mentioned that the direct age simulation can also be performed in the sense of a *backtracing*. In this case the boundary conditions for the groundwater age have to be 'reversed' (inflow boundaries becoming outflow boundaries, and *vice versa*). For backtracing a *reverse flow field* is necessary during the age computations. FEFLOW provides a specific option termed as 'Reverse flow field' which can be set in the 'Specific option settings' menu. Further modifications of Goode's method appears possible. The extensions to transient flow problems are described by Goode[5] and Varne & Carrera[8].

8.5 Concluding Remarks

The computation of the age and residence times of groundwater can be easily and efficiently performed by the present Goode method. It is applicable in both 2D and 3D cases. The direct age simulation is a welcome completion of particle tracking approaches; especially in such cases if effects of hydrodynamic dispersion becomes important, e.g., for capture zone assessments, or, more generally, in order to make cross-checks against the traditional particle tracking method and to obtained closed and better representations of residence
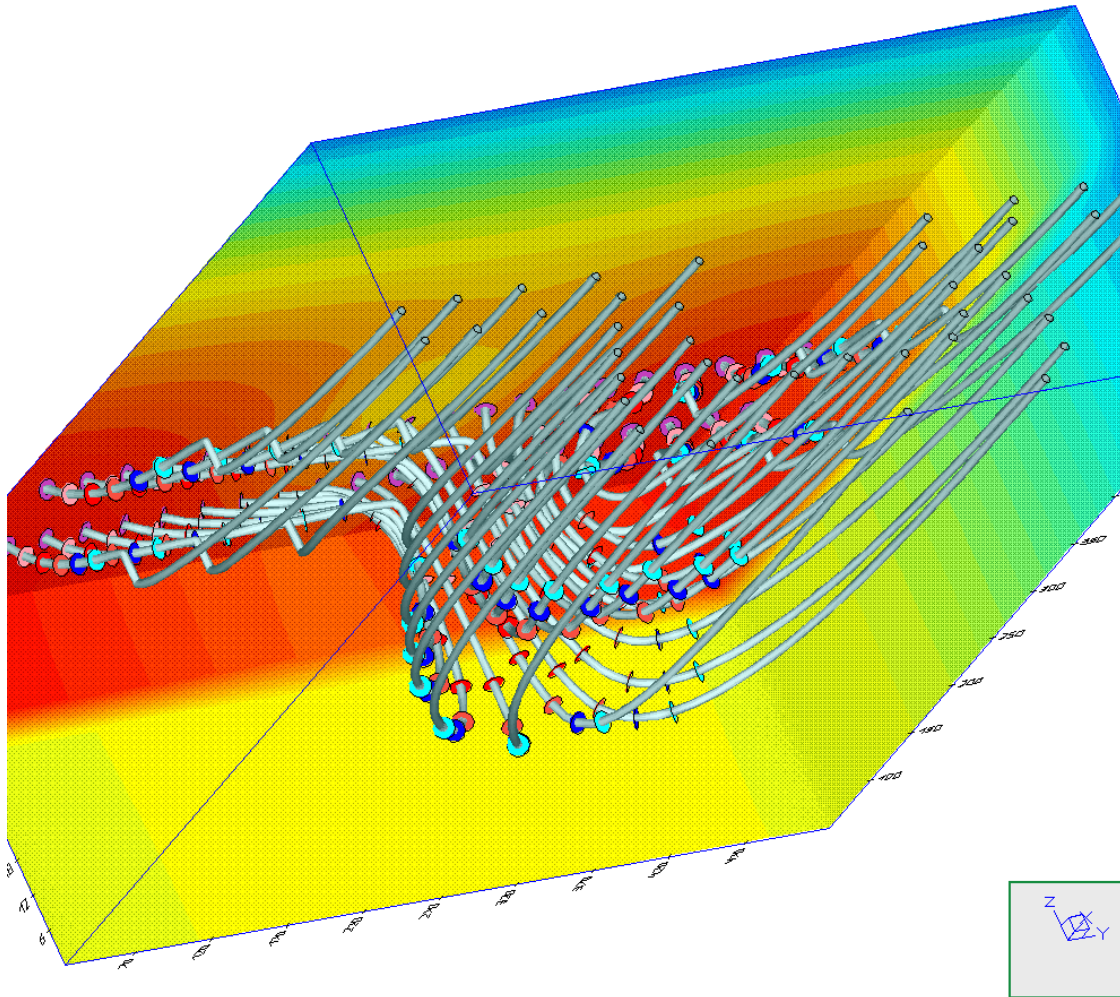
Figure 8.4 3D pathlines at selected starting points and isochrones marked at 10, 20, 30, 40 and 50 years as computed by FEFLOW's particle tracking method.
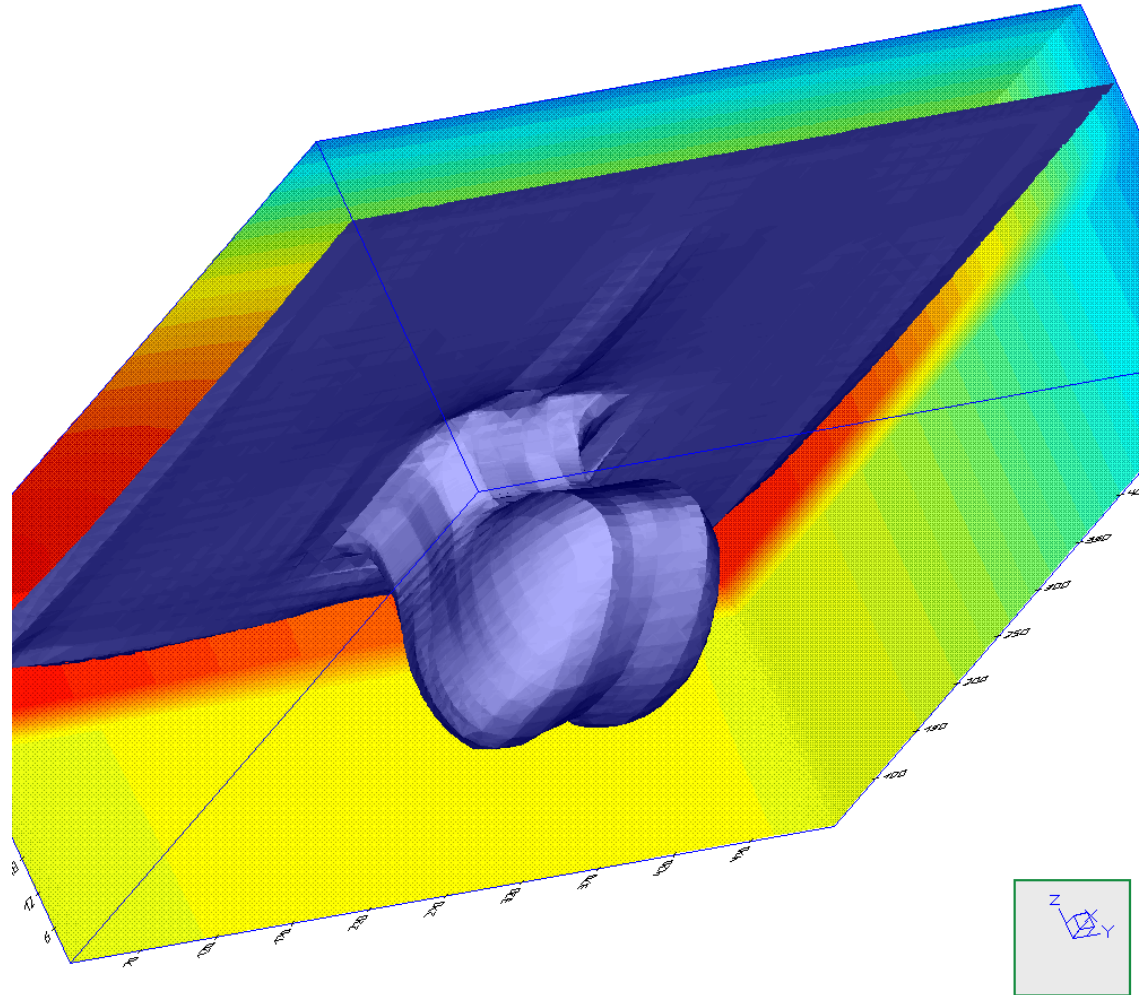
Figure 8.5 Computed age distribution for 50 years in the 3D flow domain forming a 3D isosurface.

References

1. Bear, J. & Verruijt, A., *Modeling groundwater flow and pollution.* Reidel Publ., Dordrecht, 1987.

2. Diersch, H.-J. G, Voigt, R. & Gründler, R., Visiometric techniques in a 3D groundwater transport code, In: *X Intern. Conf. on Computational Methods in Water Resources*, Heidelberg (Germany), July 19-22, Vol.2 1994, Kluwer Academic Publishers, Dordrecht, 1449-1456.

3. Diersch, H.-J. G., Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., 2002.

4. Etcheverry, D. & Perrochet, P., Direct simulation of groundwater transit-time distributions using the reservoir theory. *Hydrogeology Journal* **8** (2000) 200-208.

5. Goode, D. J., Direct simulation of groundwater age. *Water Resources Research* **32** (1996) 2, 289-296.

6. Kinzelbach, W., Numerische Methoden zur Modellierung des Transports von Schadstoffen im Grundwasser. 2nd edition, Oldenbourg Verlag, München/Wien, 1992.

7. Perrochet, P., Personal communication, Centre d'Hydrogéologie, Université de Neuchâtel, Switzerland, 1999.

8. Varne, M. & Carrera, J., Simulation of groundwater age distributions. *Water Resources Research* **34** (1998) 12, 3271-3281.

8. An efficient method for computing groundwater residence times

Discrete feature modeling of flow, mass and heat transport processes by using FEFLOW

9

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

9.1 The Discrete Feature Approach

The discrete feature approach provides the crucial link between the complex geometries for subsurface and surface continua in modeling flow, contaminant mass and heat transport processes. In this holistic approach a three-dimensional geometry of the subsurface domain (aquifer system, rock masses) in describing a porous matrix structure can be combined by interconnected one-dimensional and/or two-dimensional features as shown in Fig. 9.1. In the finite-element context the three-dimensional mesh for the porous matrix can be enriched by both 'bar' (channels, mine stopes) and areal (overland, fault) elements.

9.2 The 1D and 2D Discrete Feature Elements Used

FEFLOW[2] provides 1D and 2D discrete feature elements which can be mixed with the porous matrix elements in two and three dimensions. Different laws of

fluid motion can be defined within such discrete features, e.g., *Darcy*, *Hagen-Poiseuille* or *Manning-Strickler* laws. Both the geometric and physical characteristics of the discrete feature elements provide a large flexibility in modeling complex situations. Table 9.1 summarizes the most important characteristics and typical applications for the used 1D and 2D (as well as 3D porous media) features.

Apparently, the range of applications and the dimension of the features require an unified approach, where linear and nonlinear laws of fluid motion, porous media and free fluid flows, phreatic and non-phreatic conditions as well as spatial (3D), plane (1D, 2D) and axisymmetric (1D) geometries are embodied.

9.3 Preliminaries

9.3.1 Fundamental balance statement

The conservation of mass, momentum and energy is described by the balance statement[2] (symbols are listed

in the Appendix A 'Nomenclature')

$$\frac{\partial(\rho\psi)}{\partial t} + \nabla \cdot (\rho\psi\boldsymbol{v}) + \nabla \cdot \boldsymbol{j} = \rho f \qquad (9\text{-}1)$$

conserving the (extensive) quantity $(\rho\psi)$. Individual balance laws for $(\rho\psi)$, $\boldsymbol{j}$ and ρf are summarized in Tab. 9.2.
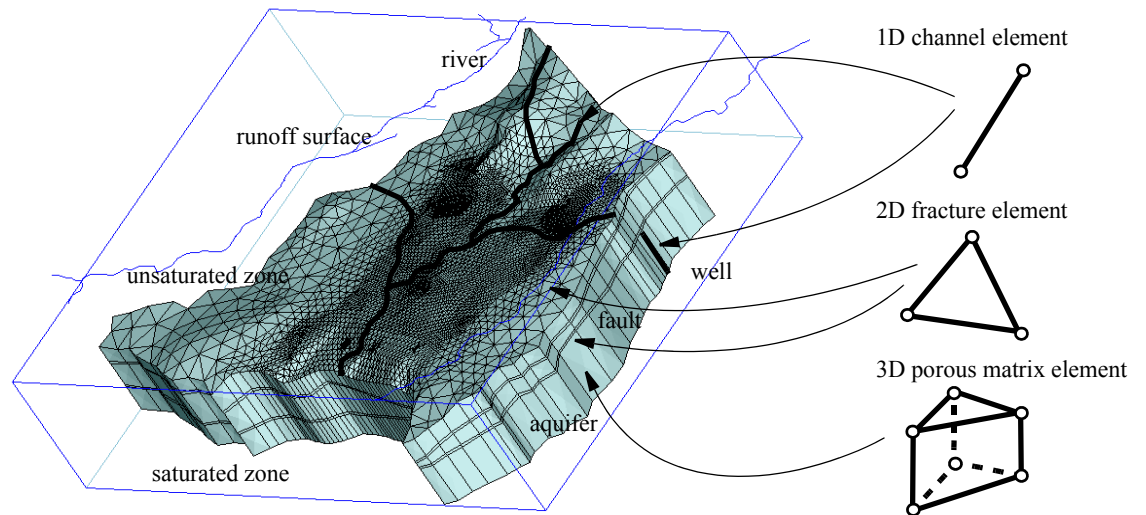

Figure 9.1 Schematization of a subsurface modeling system by combining discrete feature elements with volume discretizations of the total study domain: 1D elements are used to approximate rivers, channels, wells and specific faults, 2D feature elements are appropriate for modeling runoff processes, fractured surfaces and faulty zones, and 3D elements represents the basic tessellation of the subsurface domain consisting of an aquifer-aquitard system and involving unsaturated and saturated zones.

Table 9.1 Used discrete feature elements

Type	Fluid motion law	Dimension	Application
	Darcy Hagen-Poiseuille Manning-Strickler	1D, plane (phreatic, non-phreatic)	channels mine stopes
		1D, axisymmetric (phreatic, non-phreatic)	pumping wells abandoned wells, boreholes
	Darcy Hagen-Poiseuille Manning-Strickler	2D, plane (non-phreatic)	fractures faults
		2D, plane (phreatic)	runoff overland flow
	Darcy	3D (phreatic, non-phreatic)	porous media aquifer systems

Table 9.2 Balance laws

Quantity	$\rho\psi$	j	ρf
mass fluid mass contaminant mass	ρ C	0 j_c	ρQ_ρ r_c
momentum	ρv	σ	ρg
energy	$\rho\left(E + \frac{1}{2}v^2\right)$	$\sigma \cdot v + j_T$	$\rho(g \cdot v + Q_T)$

9.3.2 Forms of balance equations

According to the applications for the discrete feature elements indicated above we are interested in four forms of the governing balance equation (9-1):

- form A: free fluid balance law
- form B: vertically integrated free fluid balance law
- form C: porous medium balance law
- form D: vertically integrated porous medium balance law

The *form A* is already represented by Eq. (9-1). A vertical integration of (9-1) over a depth B can be rigorously performed as described in[2,3,8] leading to the *form B*:

$$\frac{\partial(B\rho\psi)}{\partial t} + \nabla \cdot (B\rho\psi v) + \nabla \cdot (Bj) = B\rho f + j_\psi^{top} - j_\psi^{bottom} \quad (9\text{-}2)$$

with the new exchange terms of the quantity ψ at the top and bottom boundaries

$$\left. \begin{aligned} j_\psi^{top} &= \frac{1}{\delta S} \int_{\delta S^{top}} n^{top} \cdot [j + \rho\psi(w - v)]dS \\ j_\psi^{bottom} &= \frac{1}{\delta S} \int_{\delta S^{bottom}} n^{bottom} \cdot [j + \rho\psi(w - v)]dS \end{aligned} \right\} \quad (9\text{-}3)$$

Notice, the balance quantities of Eq. (9-2) are now averaged over the depth B.
The transformation of the balance equation (9-1) to a porous medium is performed by a spatial averaging

procedures referred to the representative elementary volume (REV) composed by fluid and solid phases. It results finally to the *form C* of the basic balance statement[3]

$$\frac{\partial(\varepsilon\rho\psi)}{\partial t} + \nabla \cdot (\varepsilon\rho\psi\boldsymbol{v}) + \nabla \cdot (\varepsilon\boldsymbol{j}) = \varepsilon\rho f + j_\psi^{\text{interface}} \quad (9\text{-}4)$$

where an exchange term at the fluid-solid interface naturally results

$$j_\psi^{\text{interface}} = \frac{1}{\delta S} \int_{\delta S^{\text{interface}}} \boldsymbol{n}^{\text{interface}} \cdot [\boldsymbol{j} + \rho\psi(\boldsymbol{w} - \boldsymbol{v})]dS \quad (9\text{-}5)$$

Notice, the balance quantities of the porous medium conservation equation (9-4) are averaged over the REV volume.

Finally, the porous medium equation (9-4) can also be vertically integrated over the depth B, which yields *form D* of the basic balance statement as

$$\frac{\partial(B\varepsilon\rho\psi)}{\partial t} + \nabla \cdot (B\varepsilon\rho\psi\boldsymbol{v}) + \nabla \cdot (B\varepsilon\boldsymbol{j}) = \quad (9\text{-}6)$$
$$B\varepsilon\rho f + j_\psi^{\text{interface}} + j_\psi^{\text{top}} - j_\psi^{\text{bottom}}$$

It is obvious, the balance statement (9-6) of form D is the most general form which encompasses all other forms when we specify the porosity ε as

$$\varepsilon = \begin{cases} \equiv 1 & \text{free fluid flow} \\ < 1 & \text{for} \quad \text{porous media flow} \end{cases} \quad (9\text{-}7)$$

the depth B as

$$B = \begin{cases} \equiv 1 & \text{non-integrated form} \\ \text{arbitrary} & \text{for} \quad \text{vertically integrated form} \end{cases} \quad (9\text{-}8)$$

the interface exchange term $j_\psi^{\text{interface}}$ as

$$j_\psi^{\text{interface}} = \begin{cases} \equiv 0 & \text{free fluid flow} \\ \neq 0 & \text{for} \quad \text{porous media flow} \end{cases} \quad (9\text{-}9)$$

and the top and bottom exchange terms $j_\psi^{\text{top}}, j_\psi^{\text{bottom}}$ as

$$(j_\psi^{\text{top}}, j_\psi^{\text{bottom}}) = \begin{cases} \equiv 0 & \text{non-integrated form} \\ \neq 0 & \text{for} \quad \text{vertically integrated form} \end{cases} \quad (9\text{-}10)$$

9.3.3 Mathematical conventions

Both Cartesian and cylindrical coordinate systems will be employed. They are defined as

$$\boldsymbol{x} = \begin{cases} x, y, z & \text{3D} \\ x, y & \text{2D} \\ x & \text{for} \quad \text{1D} \\ r, \omega, z & \text{axisymmetry} \end{cases} \quad (9\text{-}11)$$

The velocity vector $\boldsymbol{v}$ is accordingly

$$v = \begin{cases} \begin{bmatrix} u \\ v \\ w \end{bmatrix} & \text{Cartesian coordinates} \\[20pt] \begin{bmatrix} v_r \\ v_\omega \\ v_z \end{bmatrix} & \text{cylindrical coordinates} \end{cases} \quad \text{for} \qquad (9\text{-}12)$$

The scalar product $\nabla \cdot v$ is given by

$$(\nabla \cdot v) = \begin{cases} \dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y} + \dfrac{\partial w}{\partial z} & \text{3D } (x, y, z) \text{ Cartesian} \\[12pt] \dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y} & \text{2D } (x, y) \text{ Cartesian} \\[12pt] \dfrac{\partial u}{\partial x} & \text{1D } (x) \text{ Cartesian} \\[12pt] \dfrac{1}{r}\dfrac{\partial (r v_r)}{\partial r} + \dfrac{1}{r}\dfrac{\partial v_\omega}{\partial \omega} + \dfrac{\partial v_z}{\partial z} & \text{cylindrical } (r, \omega, z) \end{cases} \qquad (9\text{-}13)$$

and the derivative operation ∇^2 for the different coordinate systems is given, for instance, for the variable ψ as

$$(\nabla^2 \psi) = \begin{cases} \dfrac{\partial^2 \psi}{\partial x^2} + \dfrac{\partial^2 \psi}{\partial y^2} + \dfrac{\partial^2 \psi}{\partial z^2} & \text{3D } (x, y, z) \text{ Cartesian} \\[12pt] \dfrac{\partial^2 \psi}{\partial x^2} + \dfrac{\partial^2 \psi}{\partial y^2} & \text{2D } (x, y) \text{ Cartesian} \\[12pt] \dfrac{\partial^2 \psi}{\partial x^2} & \text{1D } (x) \text{ Cartesian} \\[12pt] \dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial \psi}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial^2 \psi}{\partial \omega^2} + \dfrac{\partial^2 \psi}{\partial z^2} & \text{cylindrical } (r, \omega, z) \end{cases} \qquad (9\text{-}14)$$

9.3.4 Gravity and variables

In the following we assume an exclusive action of gravity in the form

$$g = -ge \qquad g = \begin{bmatrix} g_x \\ g_y \\ g_z \end{bmatrix} \qquad e = \begin{bmatrix} e_x \\ e_y \\ e_z \end{bmatrix} \qquad (9\text{-}15)$$

As a further useful variable the hydraulic head h (piezometric head) related to the reference fluid density ρ_o is defined

$$h = \phi + z = \frac{p}{\rho_o g} + z \qquad (9\text{-}16)$$

and

$$p = \rho_o g (h - z) \qquad (9\text{-}17)$$

Thus,

$$\nabla p - \rho\boldsymbol{g} = \nabla p + \rho g\boldsymbol{e} = \rho_o g\left(\nabla\phi + \nabla z + \frac{\rho - \rho_o}{\rho_o}\boldsymbol{e}\right) \quad (9\text{-}18)$$

$$= \rho_o g\left(\nabla h + \frac{\rho - \rho_o}{\rho_o}\boldsymbol{e}\right) = \rho_o g(\nabla h + \Theta\boldsymbol{e})$$

9.3.5 Hydraulic radius

The hydraulic radius is defined as the flow cross-sectional area divided by the wetted perimeter

$$r_{hydr} = \frac{\text{flow area}}{\text{wetted perimeter}} \quad (9\text{-}19)$$

Table 9.3 lists the hydraulic radii for interesting cases.

9.3.6 Free (phreatic) surface condition
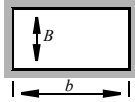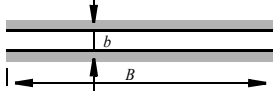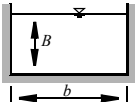
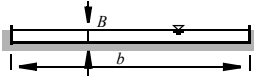A phreatic surface represents a macroscopic moving material interface between two fluids, e.g. air and water. A material surface $F = F(\boldsymbol{x}, t) = 0$ is governed by the kinematic equation

$$\frac{\partial F}{\partial t} + \boldsymbol{w} \cdot \nabla F = 0 \quad (9\text{-}20)$$

The outward unit vector normal to F is defined as

$$\boldsymbol{n} = \frac{\nabla F}{\|\nabla F\|} \quad (9\text{-}21)$$

Table 9.3 Hydraulic radii for different applications

	Type	r_{hydr}
A	submerged rectangular cross-section	$\dfrac{Bb}{2(b + B)}$
B	submerged slit plane	$\dfrac{bB}{2B} = \dfrac{b}{2}$
C	open rectangular cross-section	$\dfrac{Bb}{b + 2B}$
D	open wide channel ($b > 20B$) plane	$\dfrac{B}{1 + 2B/b} \approx B$
E	submerged circular cross-section	$\dfrac{\pi R^2}{2\pi R} = \dfrac{R}{2}$

and accordingly

$$\boldsymbol{w} \cdot \boldsymbol{n} = -\frac{\partial F/\partial t}{\|\nabla F\|} \quad (9\text{-}22)$$

where $\|\nabla F\|$ denotes the magnitude of the vector ∇F. For the vertical integration along the thickness B we can express the geometries of the top and bottom surfaces in the forms (Fig. 9.2)

$$\left.\begin{array}{l} F^{top} \equiv \quad F^{top}(x,t) = z - b^{top}(x,y,t) \quad = 0 \\ F^{bottom} \equiv F^{bottom}(x,t) = z - b^{bottom}(x,y,t) = 0 \end{array}\right\} \quad (9\text{-}23)$$

and

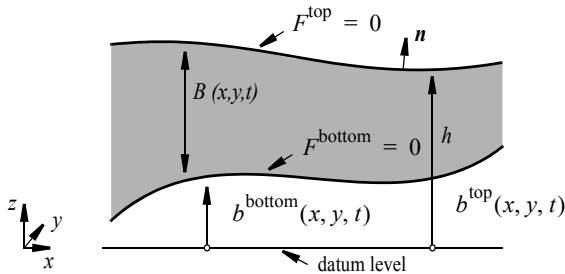$$B = B(x,t) = b^{top}(x,y,t) - b^{bottom}(x,y,t) \qquad (9\text{-}24)$$

Figure 9.2 Surface conditions.

For a free surface the top elevation $z = b^{top}(x,y,t)$ is identical to the hydraulic head $h = h(x,y,t)$. Accordingly, the thickness is given by

$$B = h - b^{bottom} \qquad (9\text{-}25)$$

If the bottom geometry is stationary the storage term in (9-6) becomes

$$\frac{\partial(B\varepsilon\rho\psi)}{\partial t} = B\frac{\partial(\varepsilon\rho\psi)}{\partial t} + \varepsilon\rho\psi\frac{\partial h}{\partial t} \qquad (9\text{-}26)$$

9.3.7 Viscous stresses on surfaces

The viscous stresses on a surface ν (note ν can indicate a *top* and *bottom* surface as well as a fluid-solid *interface*) result from exchange relationships (9-3) and (9-5) if replacing the general flux vector $\boldsymbol{j}$ by the viscous stress tensor of fluid σ (*cf.* Table 9.2), *viz.*,

$$\sigma^\nu = \frac{1}{\delta S}\int_{\delta S^\nu} \boldsymbol{n}^\nu \cdot [\sigma + \rho\boldsymbol{v}(\boldsymbol{w} - \boldsymbol{v})]dS \qquad (9\text{-}27)$$

Here σ^ν stands for the stress on the surface ν with normal $\boldsymbol{n}^\nu$. It represents a surface force per unit area depending on the orientation of the surface[10]. For instance, let us consider the stress components on a planar top surface as illustrated in Fig. 9.3. Assuming additionally a rigid and impermeable surface ($\boldsymbol{w} = \boldsymbol{v} \approx 0$) with a constant stress property on the unit area δS the surface stresses are explicitly given by

$$\sigma^{top} \approx \boldsymbol{n}^{top} \cdot \sigma \qquad (9\text{-}28)$$

With $\boldsymbol{n}^{top} = (0, 1, 0)$ the stress components become

$$\left.\begin{array}{l} \sigma_x^{top} = 0\sigma_{xx} + 1\sigma_{yx} + 0\sigma_{zx} = \sigma_{yx} \\ \sigma_y^{top} = 0\sigma_{xy} + 1\sigma_{yy} + 0\sigma_{zy} = \sigma_{yy} \\ \sigma_z^{top} = 0\sigma_{xz} + 1\sigma_{yz} + 0\sigma_{zz} = \sigma_{yz} \end{array}\right\} \qquad (9\text{-}29)$$
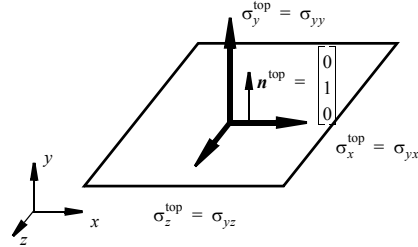
Figure 9.3 Surface forces related to components of the viscous stress tensor σ.

9.4 Basic Balance Equations

9.4.1 Fluid mass conservation

The fluid mass conservation is described by specifying Eq. (9-6) with Table 9.2 as

$$\frac{\partial}{\partial t}(\varepsilon\rho B) + \nabla \cdot (\varepsilon\rho B\boldsymbol{v}) = \varepsilon\rho B\overline{Q}_\rho \qquad (9\text{-}30)$$

which can be employed for all flow problems under discussion when setting ε and B appropriately. Notice, the sink/source term $\overline{Q}_\rho$ includes both interfacial and surfacial flux conditions (*cf.*, Eq. (9-5)).
The storage term in (9-30)

$$\frac{\partial}{\partial t}(\varepsilon\rho B) = \varepsilon B\frac{\partial\rho}{\partial t} + \rho B\frac{\partial\varepsilon}{\partial t} + \varepsilon\rho\frac{\partial B}{\partial t} \qquad (9\text{-}31)$$

can be expanded with regard to the hydraulic head h and one gets with (9-26)

$$\frac{\partial}{\partial t}(\varepsilon\rho B) = \rho(BS_o + S_s)\frac{\partial h}{\partial t} \qquad (9\text{-}32)$$

where the compressibility S_o and storativity S_s are introduced as

$$\left.\begin{array}{l} S_o = \varepsilon\gamma + (1 - \varepsilon)\kappa \\ S_s = \varepsilon \end{array}\right\} \qquad (9\text{-}33)$$

Notice, for a free fluid we have to set $\varepsilon = 1$ and $S_o = \gamma$.
Neglecting the density effects in the divergence term of (9-30) by applying the Boussinesq approximation[2] the fluid mass balance equation (9-30) yields

$$\left.\begin{array}{l} S\dfrac{\partial h}{\partial t} + \nabla \cdot (\varepsilon B\boldsymbol{v}) = \varepsilon B\overline{Q}_\rho \\ S = (BS_o + S_s) \end{array}\right\} \qquad (9\text{-}34)$$

9.4.2 Fluid momentum conservation

The fluid momentum conservation is specified from Eq. (9-6) with Table 9.2 as

$$\frac{\partial}{\partial t}(\varepsilon\rho B\boldsymbol{v}) + \nabla \cdot (\varepsilon\rho B\boldsymbol{v}\boldsymbol{v}) = -\nabla(\varepsilon Bp) \qquad (9\text{-}35)$$
$$+ \nabla \cdot (\varepsilon B\sigma') + \varepsilon\rho B\boldsymbol{g} + \varepsilon B(\sigma^{\text{interface}} + \sigma^{\text{top}} - \sigma^{\text{bottom}})$$

where the stress tensor is splitted into the equilibrium (pressure) and non-equilibrium (deviatory) parts as

$$\sigma = -p\boldsymbol{I} + \sigma' \qquad (9\text{-}36)$$

Notice, in Eq. (9-35) the exchange term $\sigma^{\text{interface}}$ vanishes for free fluid motion and the terms σ^{top}, σ^{bottom} are dropped if the equation is not vertically integrated.

In the following we assume the *Newton*'s viscosity law (including the *Stokes*'s assumption[10]) which is written in the form

$$\sigma' = 2\mu\left[d - \frac{1}{3}(\nabla \cdot v)I\right] \qquad (9\text{-}37)$$

with the strain-rate tensor

$$d = \frac{1}{2}[\nabla v + (\nabla v)^T] \qquad (9\text{-}38)$$

For an incompressible fluid with a divergenceless (so-called solenoidal) velocity $\nabla \cdot v = 0$ the momentum equation (9-35) leads to the well-known *Navier-Stokes equation*

$$\varepsilon\rho B\frac{\partial v}{\partial t} + (\varepsilon\rho Bv \cdot \nabla)v = -\varepsilon B(\nabla p - \rho g) + \varepsilon B\mu\nabla^2 v \quad (9\text{-}39)$$
$$+ \varepsilon B(\sigma^{\text{interface}} + \sigma^{\text{top}} - \sigma^{\text{bottom}})$$

from where specific forms can be derived as follows.

9.4.2.1 Darcy flow in porous media

Commonly, in a porous medium the velocity v is sufficiently small, that means the Reynolds number based on a typical pore diameter is of order unity or smaller. As the result, the inertial effects in the momentum equation (9-39) can be neglected

$$\frac{\partial v}{\partial t} \approx 0 \qquad (v \cdot \nabla)v \approx 0 \qquad (9\text{-}40)$$

As the result, one yields a general momentum equation for porous media (we consider the non-integrated form with $B \equiv 1$, $\sigma^{\text{top}} = \sigma^{\text{bottom}} = 0$) as

$$\varepsilon(\nabla p - \rho g) = \varepsilon\sigma^{\text{interface}} + \varepsilon\mu\nabla^2 v \qquad (9\text{-}41)$$

Furthermore, the drag forces due to fluid viscosity can usually be dropped $\mu\nabla^2 v \approx 0$ with respect to the drag term of momentum exchange $\sigma^{\text{interface}}$ at interfaces of phases. The interfacial drag term of momentum exchange $\sigma^{\text{interface}}$ can be derived as a linear friction relationship of the form[3]

$$\sigma^{\text{interface}} = -\mu k^{-1} \cdot (\varepsilon v) \qquad (9\text{-}42)$$

where the permeability k represents an inverse friction tensor due to the viscous drag at the interfaces of fluid-solid phases.

Finally, the momentum equation (9-41) reduces to the well-known *Darcy equation* of the form

$$v = -\frac{k}{\varepsilon\mu}(\nabla p - \rho g) \qquad (9\text{-}43a)$$

or with (9-18)

$$\left.\begin{array}{l} \varepsilon v = -Kf_\mu(\nabla h + \Theta e) \\[4pt] K = \dfrac{k\rho_o g}{\mu_o} \\[4pt] f_\mu = \dfrac{\mu_o}{\mu} \end{array}\right\} \qquad (9\text{-}43b)$$

valid for flow in a porous medium.

$$\frac{dp}{dx} - \rho g_x = \mu \frac{d^2 u}{dy^2} \tag{9-45}$$

9.4.2.2 Plane and axisymmetric parallel (Poiseuille) flow

A flow is called parallel when inertial terms of the Navier-Stokes equation (9-39) vanishes. That means, a fluid particle is subjected to zero acceleration, accordingly, it moves in pure translation with constant velocity. It follows that pathlines must be straight lines and that the velocity of each particle may depend only on coordinates perpendicular to the direction of flow. Such flow fields occur between two parallel plates or in a circular tube as depicted in Fig. 9.4.
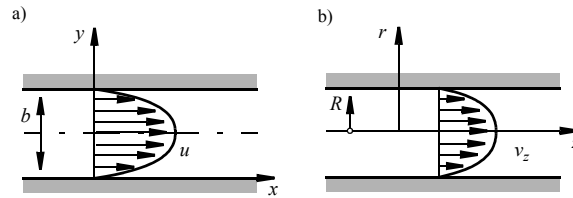
Integrating (9-45) with the boundary conditions $u(0) = u(b) = 0$ it yields

$$u = -\frac{1}{2\mu}\left(\frac{dp}{dx} - \rho g_x\right)y(b-y) \tag{9-46}$$

and we obtain the average velocity in the aperture b as

$$\bar{u} = \frac{1}{b} \int\limits_{y=0}^{b} u\, dy = -\frac{b^2}{12\mu}\left(\frac{dp}{dx} - \rho g_x\right) \tag{9-47}$$

and the discharge Q

$$Q = \bar{u} b = -\frac{b^3}{12\mu}\left(\frac{dp}{dx} - \rho g_x\right) \tag{9-48}$$

which is called the *cubic law* of the Hagen-Poiseuille flow. The relationships (9-47) can be expressed by the hydraulic radius r_{hydr} if replacing the dimension $b/2$ for the slit flow according to Table 9.3 (type B)

Figure 9.4 a) 2D plane and b) axisymmetric Poiseuille flow.

For 2D parallel laminar flow (Fig. 9.4a) we have

$$\mathbf{v} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \qquad u = u(y) \qquad v = w = 0 \tag{9-44}$$

and the momentum equation (9-39) in the *x*-direction becomes (we consider the free fluid case with no vertical integration)

$$\bar{u} = -\frac{r_{hydr}^2}{3\mu}\left(\frac{dp}{dx} - \rho g_x\right) \tag{9-49}$$

Similarly, for the axisymmetric flow in a circular tube (Fig. 9.4b) with

$$v = \begin{bmatrix} v_r \\ v_\omega \\ v_z \end{bmatrix} \qquad v_z = v_z(r) \qquad v_r = v_\omega = 0 \qquad (9\text{-}50)$$

one solves in the z-direction the momentum equation

$$\frac{dp}{dz} - \rho g_z = \frac{\mu}{r}\left[\frac{\partial}{\partial r}\left(r\frac{\partial v_z}{\partial r}\right)\right] \qquad (9\text{-}51)$$

With $dv_z/dr = 0$ at $r = 0$ and $v_z(R) = 0$ the integration of (9-51) gives

$$v_z = -\frac{1}{4\mu}\left(\frac{dp}{dz} - \rho g_z\right)(R^2 - r^2) \qquad (9\text{-}52)$$

Then, the average velocity for the Hagen-Poiseuille flow in a circular tube becomes

$$\bar{v}_z = \frac{1}{\pi R^2}\int_{\omega=0}^{2\pi}\int_{r=0}^{R} v_z r\, dr\, d\omega = -\frac{R^2}{8\mu}\left(\frac{dp}{dz} - \rho g_z\right) \qquad (9\text{-}53)$$

and we get for the discharge through the tube

$$Q = \pi R^2 \bar{v}_z = -\frac{\pi R^4}{8\mu}\left(\frac{dp}{dz} - \rho g_z\right) \qquad (9\text{-}54)$$

The relationships (9-53) can be expressed by the hydraulic radius r_{hydr} if replacing the dimension $R/2$ for the tube flow according to Table 9.3 (type E)

$$\bar{v}_z = -\frac{r_{\text{hydr}}^2}{2\mu}\left(\frac{dp}{dz} - \rho g_z\right) \qquad (9\text{-}55)$$

As seen the Hagen-Poiseuille's laws of laminar fluid motion for 1D and 2D plane flow (9-47) and for axisymmetric flow (9-53) represent linear relationships with respect to the pressure gradient and gravity $(\nabla p - \rho g)$. In a generalized form one yields finally with (9-18)

$$v = -K f_\mu(\nabla h + \Theta e) \qquad (9\text{-}56)$$

$$K = \frac{r_{\text{hydr}}^2 \rho_o g}{a\mu_o} I \quad \text{with} \begin{cases} r_{\text{hydr}} = b/2, a = 3 & \text{for 1D/2D plane} \\ r_{\text{hydr}} = R/2, a = 2 & \text{for axisymmetry} \end{cases}$$

9.4.2.3 Laws of fluid motion for overland and channel flow

Basically, the fluid motion for overland and channel flow is described by the vertically integrated Navier-Stokes equation (9-39) according to

$$\rho B\frac{\partial v}{\partial t} + (\rho B v \cdot \nabla)v = -B(\nabla p - \rho g) \qquad (9\text{-}57)$$
$$+ B\mu\nabla^2 v + B(\sigma^{\text{top}} - \sigma^{\text{bottom}})$$

which is a formulation of the well-known *De Saint-Venant* equations[1]. Over a wide range of practical overland and channel flow (Fig. 9.5) at low-to-moderate velocity/flow regimes the inertial terms in the governing momentum balance equation (9-35) can be ignored compared with the gravitational terms, friction and pressure effects. Furthermore, the interior viscous effects can be neglected over the shear stress effects at the surfaces[1,9]. Assuming this,

$$\frac{\partial \boldsymbol{v}}{\partial t} \approx 0 \qquad (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} \approx 0 \qquad \mu\nabla^2 \boldsymbol{v} \approx 0 \qquad (9\text{-}58)$$

the momentum equation (9-57) reduces to

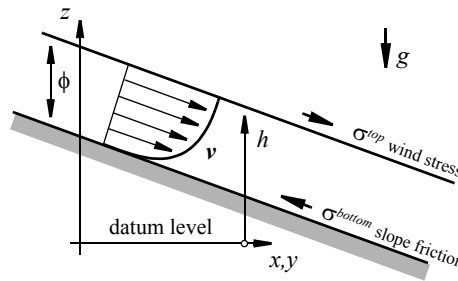$$(\nabla p - \rho \boldsymbol{g}) - \sigma^{top} + \sigma^{bottom} = 0 \qquad (9\text{-}59)$$

Figure 9.5 Open channel flow.

The shear effect σ^{top} at the top (free) surface can be caused by wind stress. For the present application we neglect influences caused by wind stress:

$$\sigma^{top} \approx 0 \qquad (9\text{-}60)$$

On the other hand, the shear effect at the bottom can be expressed by a friction slope relationship of the form

$$\sigma^{bottom} = \frac{\rho_o g \|\boldsymbol{v}\| \boldsymbol{v}}{\tau^2 r_{hydr}^{\alpha}} \qquad (9\text{-}61)$$

representing friction laws, where $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}}$, τ is a friction factor and $\alpha \geq 1$ is a constant. As the result, the

following momentum equation can be derived

$$\left.\begin{array}{c}(\nabla p - \rho \boldsymbol{g}) + \rho_o g \boldsymbol{S}_f = 0 \\[2mm] \boldsymbol{S}_f = \dfrac{\|\boldsymbol{v}\| \boldsymbol{v}}{\tau^2 r_{hydr}^{\alpha}}\end{array}\right\} \qquad (9\text{-}62)$$

where different specific laws for the friction slopes $\boldsymbol{S}_f$ can be specified as summarized in Table 9.4 for isotropic roughness coefficients.

Table 9.4 Various friction laws

Law	τ	α	$\boldsymbol{S}_f$
Newton-Taylor	$\sqrt{\dfrac{g}{\Upsilon}}$	1	$\dfrac{\Upsilon \|\boldsymbol{v}\| \boldsymbol{v}}{g r_{hydr}}$
Chezy	C	1	$\dfrac{\|\boldsymbol{v}\| \boldsymbol{v}}{C^2 r_{hydr}}$
Manning-Strickler	M	$4/3$	$\dfrac{\|\boldsymbol{v}\| \boldsymbol{v}}{M^2 r_{hydr}^{4/3}}$

Instead of using the pressure p as primary variable the hydraulic head h or the local water depth ϕ (*cf.* Eqs. (9-16) and (9-18)) are alternative formulations of (9-62), *viz.*,

$$\rho_o g(\nabla h + \boldsymbol{S}_f + \Theta \boldsymbol{e}) = 0 \qquad (9\text{-}63a)$$

and

$$\rho_o g(\nabla \phi + \boldsymbol{S}_f - \boldsymbol{S}_o + \Theta \boldsymbol{e}) = 0 \left.\right\}$$
$$\boldsymbol{S}_o = -\nabla z$$

(9-63b)

where fluid density effects are included in the $\Theta \boldsymbol{e}$-term.

Equation (9-63a) can be used to derive a *diffusion-type flow equation*[7,9]. Since, exemplified for 2D

$$\|\boldsymbol{v}\|^2 = u^2 + v^2 = \tau^2 r_{hydr}^\alpha \sqrt{S_{fx}^2 + S_{fy}^2}$$

(9-64)

and with (9-62)

$$S_{fx} = \frac{\sqrt{u^2 + v^2}}{\tau^2 r_{hydr}^\alpha} u \qquad S_{fy} = \frac{\sqrt{u^2 + v^2}}{\tau^2 r_{hydr}^\alpha} v$$

(9-65)

we find with (9-63a): $S_{fx} = -(\partial h / \partial x + \Theta e_x)$, $S_{fy} = -(\partial h / \partial y + \Theta e_y)$

$$u = -\frac{\tau r_{hydr}^{\alpha/2}}{4\sqrt{\left(\frac{\partial h}{\partial x}\right)^2 + \left(\frac{\partial h}{\partial y}\right)^2}} \left(\frac{\partial h}{\partial x} + \Theta e_x\right)$$

(9-66)

$$v = -\frac{\tau r_{hydr}^{\alpha/2}}{4\sqrt{\left(\frac{\partial h}{\partial x}\right)^2 + \left(\frac{\partial h}{\partial y}\right)^2}} \left(\frac{\partial h}{\partial y} + \Theta e_y\right)$$

or more general

$$\boldsymbol{v} = -\boldsymbol{K}(\nabla h + \Theta \boldsymbol{e}) \left.\right\}$$
$$\boldsymbol{K} = \frac{\tau r_{hydr}^{\alpha/2}}{4\sqrt{\|\nabla h\|^2}} \boldsymbol{I}$$

(9-67)

It can be easily shown that the velocity $\boldsymbol{v}$ in the relationship (9-67) tends to zero if the gradient ∇h vanishes (provided that $\Theta \approx 0$)

$$\lim_{\nabla h \to 0} \boldsymbol{v} = -\lim_{\nabla h \to 0} \frac{\tau r_{hydr}^{\alpha/2}}{4\sqrt{\|\nabla h\|^2}} \boldsymbol{I} \nabla h = 0$$

(9-68)

9.4.3 Contaminant mass conservation

The balance equation for a contaminant mass results from Eq. (9-6) and Table 9.2 in form of

$$\frac{\partial(B\varepsilon C)}{\partial t} + \nabla \cdot (B\varepsilon C\boldsymbol{v}) + \nabla \cdot (B\varepsilon \boldsymbol{j}_c) = B\varepsilon \bar{r}_c$$

(9-69)

which can be employed for all the interesting mass transport problems when specifying ε and B appropriately. Notice, the reaction term $\bar{r}_c$ includes both interfacial and surfacial mass transfer conditions (*cf.*, Eq. (9-5)).

The reaction term can be splitted into a first-order reaction rate and a zero-order production term[2], respectively,

$$\bar{r}_c = -\vartheta C + Q_c$$

(9-70)

The mass flux $\boldsymbol{j}_c$ is expressed by the *Fickian law* in form of

$$\boldsymbol{j}_c = -\boldsymbol{D} \cdot \nabla C \left.\right\}$$
$$\boldsymbol{D} = D_d \boldsymbol{I} + \boldsymbol{D}_m$$

(9-71)

The hydrodynamic dispersion tensor $\boldsymbol{D}$ consists of the molecular diffusion part $D_d\boldsymbol{I}$ and the mechanical dispersion part $\boldsymbol{D}_m$. In a porous medium $\boldsymbol{D}_m$ is commonly described by the Scheidegger-Bear dispersion relationship as

$$\boldsymbol{D}_m = (\beta_T\|\boldsymbol{v}\|)\boldsymbol{I} + (\beta_L - \beta_T)\frac{\boldsymbol{v} \otimes \boldsymbol{v}}{\|\boldsymbol{v}\|} \qquad (9\text{-}72)$$

In a free fluid flow there is a large variety for $\boldsymbol{D}_m$ in dependence on laminar and turbulent flow conditions. For instance, in a fluid-filled tube under laminar flow conditions $\boldsymbol{D}_m$ can by estimated by Taylor's analysis[11]

$$\boldsymbol{D}_m = \left(\frac{R^2\|\boldsymbol{v}\|}{48 D_d}\right)\frac{\boldsymbol{v} \otimes \boldsymbol{v}}{\|\boldsymbol{v}\|} \qquad (9\text{-}73)$$

Using the Fickian law (9-71) and incorporating the continuity equation (9-30) Eq. (9-69) yields[2]

$$\varepsilon B\frac{\partial C}{\partial t} + \varepsilon B\boldsymbol{v} \cdot \nabla C - \nabla \cdot (B\varepsilon\boldsymbol{D} \cdot \nabla C) \qquad (9\text{-}74)$$
$$+ B\varepsilon(\overline{Q}_\rho + \vartheta)C = B\varepsilon Q_c$$

Considering additionally sorption effects in the porous medium the following contaminant mass transport equation can be derived on the basis of Eq. (9-74)[2]

$$B\Re_d\frac{\partial C}{\partial t} + \varepsilon B\boldsymbol{v} \cdot \nabla C - \nabla \cdot (B\varepsilon\boldsymbol{D} \cdot \nabla C) \qquad (9\text{-}75)$$
$$+ B(\varepsilon\overline{Q}_\rho + \Re\vartheta)C = B\varepsilon Q_c$$

with the retardation relationships

$$\left.\begin{aligned}\Re &= \varepsilon + (1-\varepsilon)\chi(C) \\ \Re_d &= \varepsilon + (1-\varepsilon)\frac{d}{dC}[\chi(C) \cdot C]\end{aligned}\right\} \qquad (9\text{-}76)$$

in which the sorption function $\chi(C)$ can be specified for Henry, Freundlich or Langmuir isotherms[2].

9.4.4 Energy conservation

The energy balance equation is derived basically from Eq. (9-6) and Table 9.2 under the assumption of a thermal equilibrium between fluid (f) and solid (s) phases. We obtain finally[2]

$$\frac{\partial}{\partial t}\{B[\varepsilon\rho^f E^f + (1-\varepsilon)\rho^s E^s]\} + \nabla \cdot (B\varepsilon\rho^f E^f \boldsymbol{v}) \qquad (9\text{-}77)$$
$$+ \nabla \cdot (B\boldsymbol{j}_T) = B[\varepsilon\rho^f \overline{Q}_T^f + (1-\varepsilon)\rho^s \overline{Q}_T^s]$$

which can be applied to all the interesting heat transport problems when specifying ε and B appropriately. Notice, the thermal sink/source terms $\overline{Q}_T^f, \overline{Q}_T^s$ include both interfacial and surfacial heat transfer conditions (*cf.*, Eq. (9-5)).

Using the state relation for the internal energy[2]

$$dE^\alpha = c^\alpha dT \qquad \text{for} \qquad \alpha = s, f \qquad (9\text{-}78)$$

and the *Fourierian heat flux* as

$$\boldsymbol{j}_T = -\Lambda \cdot \nabla T$$

$$\left. \Lambda = \Lambda^{cond} + \Lambda^{disp} = [\varepsilon \lambda^f + (1-\varepsilon)\lambda^s]\boldsymbol{I} + \varepsilon \rho^f c^f \boldsymbol{D}_m \right\} \text{(9-79)}$$

one yields the following balance equation for the thermal energy[2]

$$\{B[\varepsilon \rho^f c^f + (1-\varepsilon)\rho^s c^s]\}\frac{\partial T}{\partial t} + \varepsilon \rho^f c^f B \boldsymbol{v} \cdot \nabla T$$

$$-\nabla \cdot (B\Lambda \cdot \nabla T) + B\varepsilon \rho^f c^f \overline{Q}_\rho (T - T_o) \qquad \text{(9-80)}$$

$$= B[\varepsilon \rho^f \overline{Q}_T^f + (1-\varepsilon)\rho^s \overline{Q}_T^s]$$

to be solved for the system temperature T.

9.5 Generalized Model Equations

9.5.1 Flow

The fundamental flow equation represents a combination of the fluid mass conservation equation (9-34) and the fluid momentum conservations for porous media (9-43b), Poiseuille flow (9-56) and overland/channel flow (9-67). As the result, Table 9.5 summarizes the governing equation for the used discrete feature elements for 1D, 2D and 3D in dependence on the problem cases under consideration. For the Poiseuille flow and overland/channel flow standard geometric forms of the fractures are implemented in FEFLOW. Different geometries can be input by means of corrections in the corresponding hydraulic parameters as thoroughly described in Appendix D.

9.5.2 Contaminant mass

The governing contaminant mass transport equation (9-75) can now be specified for the different flow conditions and discrete feature elements. Table 9.6 summarizes the different terms and expressions for both porous media and free fluid conditions.

9.5.3 Heat

The specified terms for the governing heat transport equation (9-80) are summarized in Table 9.7 for both porous media and free fluid conditions.

Table 9.5 Flow model equations

$$L(h) = S\frac{\partial h}{\partial t} - \nabla \cdot (Kf_\mu B \cdot (\nabla h + \Theta e)) - Q = 0$$

Case	S Darcy	S Poiseuille	S Overland	$Kf_\mu B$ Darcy	$Kf_\mu B$ Poiseuille	$Kf_\mu B$ Overland	Q Darcy	Q Poiseuille	Q Overland
1DPP (1D plane phreatic)	$b(BS_o + S_s)$	$b(B\gamma + 1)$	$b(B\gamma + 1)$	$bB \underbrace{\frac{\boldsymbol{k}\rho_o g}{\mu_o}}_{\boldsymbol{K}} f_\mu$	$bB \underbrace{\frac{r_{hydr}^2 \rho_o g \boldsymbol{I}}{3\mu_o}}_{\boldsymbol{K}} f_\mu$	$bB \underbrace{\frac{\tau r_{hydr}^{\alpha/2}\boldsymbol{I}}{\sqrt[4]{\|\nabla h\|^2}}}_{\boldsymbol{K}} f_\mu$	$bB\varepsilon\bar{Q}_\rho$	bBQ_ρ	bBQ_ρ
1DPN (1D plane non-phreatic)	bBS_o	$bB\gamma$	$bB\gamma$	$bB \underbrace{\frac{\boldsymbol{k}\rho_o g}{\mu_o}}_{\boldsymbol{K}} f_\mu$	$bB \underbrace{\frac{r_{hydr}^2 \rho_o g \boldsymbol{I}}{3\mu_o}}_{\boldsymbol{K}} f_\mu$	$bB \underbrace{\frac{\tau r_{hydr}^{\alpha/2}\boldsymbol{I}}{\sqrt[4]{\|\nabla h\|^2}}}_{\boldsymbol{K}} f_\mu$	$bB\varepsilon\bar{Q}_\rho$	bBQ_ρ	bBQ_ρ
1DAP (1D axisymmetric phreatic)	$\pi R^2\left(S_o + \frac{S_s}{B}\right)$	$\pi R^2\left(\gamma + \frac{1}{B}\right)$	$\pi R^2\left(\gamma + \frac{1}{B}\right)$	$\pi R^2 \underbrace{\frac{\boldsymbol{k}\rho_o g}{\mu_o}}_{\boldsymbol{K}} f_\mu$	$\pi R^2 \underbrace{\frac{r_{hydr}^2 \rho_o g \boldsymbol{I}}{2\mu_o}}_{\boldsymbol{K}} f_\mu$	$\pi R^2 \underbrace{\frac{\tau r_{hydr}^{\alpha/2}\boldsymbol{I}}{\sqrt[4]{\|\nabla h\|^2}}}_{\boldsymbol{K}} f_\mu$	$\pi R^2 \varepsilon\bar{Q}_\rho$	$\pi R^2 Q_\rho$	$\pi R^2 Q_\rho$
1DAN (1D axisymmetric non-phreatic)	$\pi R^2 S_o$	$\pi R^2 \gamma$	$\pi R^2 \gamma$	$\pi R^2 \underbrace{\frac{\boldsymbol{k}\rho_o g}{\mu_o}}_{\boldsymbol{K}} f_\mu$	$\pi R^2 \underbrace{\frac{r_{hydr}^2 \rho_o g \boldsymbol{I}}{2\mu_o}}_{\boldsymbol{K}} f_\mu$	$\pi R^2 \underbrace{\frac{\tau r_{hydr}^{\alpha/2}\boldsymbol{I}}{\sqrt[4]{\|\nabla h\|^2}}}_{\boldsymbol{K}} f_\mu$	$\pi R^2 \varepsilon\bar{Q}_\rho$	$\pi R^2 Q_\rho$	$\pi R^2 Q_\rho$
2DPP (2D plane phreatic)	$BS_o + S_s$	$B\gamma + 1$	$B\gamma + 1$	$B \underbrace{\frac{\boldsymbol{k}\rho_o g}{\mu_o}}_{\boldsymbol{K}} f_\mu$	$B \underbrace{\frac{r_{hydr}^2 \rho_o g \boldsymbol{I}}{3\mu_o}}_{\boldsymbol{K}} f_\mu$	$B \underbrace{\frac{\tau r_{hydr}^{\alpha/2}\boldsymbol{I}}{\sqrt[4]{\|\nabla h\|^2}}}_{\boldsymbol{K}} f_\mu$	$B\varepsilon\bar{Q}_\rho$	BQ_ρ	BQ_ρ
2DPN (2D plane non-phreatic)	BS_o	$B\gamma$	$B\gamma$	$B \underbrace{\frac{\boldsymbol{k}\rho_o g}{\mu_o}}_{\boldsymbol{K}} f_\mu$	$B \underbrace{\frac{r_{hydr}^2 \rho_o g \boldsymbol{I}}{3\mu_o}}_{\boldsymbol{K}} f_\mu$	$B \underbrace{\frac{\tau r_{hydr}^{\alpha/2}\boldsymbol{I}}{\sqrt[4]{\|\nabla h\|^2}}}_{\boldsymbol{K}} f_\mu$	$B\varepsilon\bar{Q}_\rho$	BQ_ρ	BQ_ρ
3DP (3D phreatic and non-phreatic)	S_o	γ	γ	$\underbrace{\frac{\boldsymbol{k}\rho_o g}{\mu_o}}_{\boldsymbol{K}} f_\mu$	$\underbrace{\frac{r_{hydr}^2 \rho_o g \boldsymbol{I}}{3\mu_o}}_{\boldsymbol{K}} f_\mu$	$\underbrace{\frac{\tau r_{hydr}^{\alpha/2}\boldsymbol{I}}{\sqrt[4]{\|\nabla h\|^2}}}_{\boldsymbol{K}} f_\mu$	$\varepsilon\bar{Q}_\rho$	Q_ρ	Q_ρ

Table 9.6 Contaminant mass transport model equations

$$L(C) = S\frac{\partial C}{\partial t} + q \cdot \nabla C - \nabla \cdot (B\varepsilon D \cdot \nabla C) + \Phi C - Q = 0$$

	Case	S		q		$B\varepsilon D$		Φ		Q	
		porous	free fluid	porous	free fluid	porous	free fluid	porous	free fluid	porous	free fluid
	1DPP (1D plane phreatic and non-phreatic)	$bB\Re_d$	bB	$bB\varepsilon v$	bBv	$bB\varepsilon(D_d I + D_m)$	$bB(D_d I + D_m)$	$bB(\varepsilon\bar{Q}_\rho + \Re\vartheta)$	$bB(Q_\rho + \vartheta)$	$bB\varepsilon Q_c$	bBQ_c
	1DAP (1D axisymmetric phreatic and non-phreatic)	$\pi R^2\Re_d$	πR^2	$\pi R^2\varepsilon v$	$\pi R^2 v$	$\pi R^2\varepsilon(D_d I + D_m)$	$\pi R^2(D_d I + D_m)$	$\pi R^2(\varepsilon\bar{Q}_\rho + \Re\vartheta)$	$\pi R^2(Q_\rho + \vartheta)$	$\pi R^2\varepsilon Q_c$	$\pi R^2 Q_c$
	2DPP (2D plane phreatic and non-phreatic)	$B\Re_d$	B	$B\varepsilon v$	Bv	$B\varepsilon(D_d I + D_m)$	$B(D_d I + D_m)$	$B(\varepsilon\bar{Q}_\rho + \Re\vartheta)$	$B(Q_\rho + \vartheta)$	$B\varepsilon Q_c$	BQ_c
	3DP (3D phreatic and non-phreatic)	$\Re_d$	1	εv	v	$\varepsilon(D_d I + D_m)$	$D_d I + D_m$	$\varepsilon\bar{Q}_\rho + \Re\vartheta$	$Q_\rho + \vartheta$	εQ_c	Q_c

Table 9.7 Heat transport model equations

$$L(T) = S\frac{\partial T}{\partial t} + \mathbf{q} \cdot \nabla T - \nabla \cdot (B\Lambda \cdot \nabla T) + \Phi(T - T_o) - Q = 0$$

	Case	S porous	S free fluid	q porous	q free fluid	BΛ porous	BΛ free fluid	Φ porous	Φ free fluid	Q porous	Q free fluid
	1DPP (1D plane phreatic and non-phreatic)	$bB[\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s]$	$bB\rho^f c^f$	$bB\varepsilon\rho^f c^f \mathbf{v}$	$bB\rho^f c^f \mathbf{v}$	$bB\{[\varepsilon\lambda^f + (1-\varepsilon)\lambda^s]\mathbf{I} + \varepsilon\rho^f c^f \mathbf{D}_m\}$	$bB(\lambda^f \mathbf{I} + \rho^f c^f \mathbf{D}_m)$	$bB\varepsilon\rho^f c^f \overline{Q}_\rho$	$bB\rho^f c^f Q_\rho$	$bB[\varepsilon\rho^f \overline{Q}_T^f + (1-\varepsilon)\rho^s \overline{Q}_T^s]$	$bB\rho^f Q_T^f$
	1DAP (1D axisymmetric phreatic and non-phreatic)	$\pi R^2[\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s]$	$\pi R^2 \rho^f c^f$	$\pi R^2\varepsilon\rho^f c^f \mathbf{v}$	$\pi R^2\rho^f c^f \mathbf{v}$	$\pi R^2\{[\varepsilon\lambda^f + (1-\varepsilon)\lambda^s]\mathbf{I} + \varepsilon\rho^f c^f \mathbf{D}_m\}$	$\pi R^2(\lambda^f \mathbf{I} + \rho^f c^f \mathbf{D}_m)$	$\pi R^2\varepsilon\rho^f c^f \overline{Q}_\rho$	$\pi R^2\rho^f c^f Q_\rho$	$\pi R^2[\varepsilon\rho^f \overline{Q}_T^f + (1-\varepsilon)\rho^s \overline{Q}_T^s]$	$\pi R^2\rho^f Q_T^f$
	2DPP (2D plane phreatic and non-phreatic)	$B[\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s]$	$B\rho^f c^f$	$B\varepsilon\rho^f c^f \mathbf{v}$	$B\rho^f c^f \mathbf{v}$	$B\{[\varepsilon\lambda^f + (1-\varepsilon)\lambda^s]\mathbf{I} + \varepsilon\rho^f c^f \mathbf{D}_m\}$	$B(\lambda^f \mathbf{I} + \rho^f c^f \mathbf{D}_m)$	$B\varepsilon\rho^f c^f \overline{Q}_\rho$	$B\rho^f c^f Q_\rho$	$B[\varepsilon\rho^f \overline{Q}_T^f + (1-\varepsilon)\rho^s \overline{Q}_T^s]$	$B\rho^f Q_T^f$
	3DP (3D phreatic and non-phreatic)	$\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s$	$\rho^f c^f$	$\varepsilon\rho^f c^f \mathbf{v}$	$\rho^f c^f \mathbf{v}$	$[\varepsilon\lambda^f + (1-\varepsilon)\lambda^s]\mathbf{I} + \varepsilon\rho^f c^f \mathbf{D}_m$	$\lambda^f \mathbf{I} + \rho^f c^f \mathbf{D}_m$	$\varepsilon\rho^f c^f \overline{Q}_\rho$	$\rho^f c^f Q_\rho$	$\varepsilon\rho^f \overline{Q}_T^f + (1-\varepsilon)\rho^s \overline{Q}_T^s$	$\rho^f Q_T^f$

9.6 Finite Element Formulations

9.6.1 Master equation, boundary conditions and weak statement

The governing balance equations as listed in Tables 9.5, 9.6 and 9.7 can be generalized by the following master equation

$$L(\psi) = S\frac{\partial \psi}{\partial t} + \boldsymbol{q} \cdot \nabla\psi - \nabla \cdot (\boldsymbol{D} \cdot \nabla\psi) - Q_\psi = 0$$
$$Q_\psi = -\Phi\psi + Q$$
(9-81)

which has to be solved for flow ($\psi = h$), contaminant mass ($\psi = C$) and heat ($\psi = T$) for 1D, 2D and 3D discrete feature elements.

Let $\Omega \subset R^D$ and $(0, T_t)$ be the spatial and temporal domain, respectively, where D is the number of space dimension (1, 2 or 3) and T_t is the final time, and let $\partial\Omega = \Gamma_1 \otimes \Gamma_2$ denote the boundary of Ω, where Γ_1 and Γ_2 are two disjoint portions of the total boundary, $\partial\Omega$, the following boundary conditions (B.C.'s) have to be appended to (9-81):

$$\left.\begin{array}{rll} \psi = \psi_1 & \text{on} & \Gamma_1 \\ -\boldsymbol{n} \cdot (\boldsymbol{D} \cdot \nabla\psi) + a(\psi_2 - \psi) = b & \text{on} & \Gamma_2 \end{array}\right\}$$
(9-82)

where on Γ_1 we have Dirichlet BC and on Γ_2 it represents a more general form of a Robin type BC in which more specific Neumann and Cauchy type BC's are involved. If $a = 0$ a Neumann BC of 2nd kind results, while for $b = 0$ a common Cauchy BC of 3rd kind is given. In (9-82) $\boldsymbol{n}$ corresponds to the normal unit vector (positive outward), ψ_1 and ψ_2 are prescribed boundary values of ψ on Γ_1 and Γ_2, respectively.

The finite element formulation is based on the weak form of the basic equation (9-81). Introducing a spatial weighting function w we get

$$\int_\Omega w\left(S\frac{\partial \psi}{\partial t} + \boldsymbol{q} \cdot \nabla\psi\right)d\Omega$$
(9-83)
$$= \int_\Omega w[\nabla \cdot (\boldsymbol{D} \cdot \nabla\psi) + Q_\psi]d\Omega$$

Applying partial integration and the divergence theorem (Green's theorem) to the weak statement (9-83) and inserting the Robin-type BC (9-82) the following weak form for the finite element method finally results

$$\int_\Omega \left[w\left(S\frac{\partial \psi}{\partial t} + \boldsymbol{q} \cdot \nabla\psi\right) + \nabla w \cdot (\boldsymbol{D} \cdot \nabla\psi)\right]d\Omega$$
(9-84)
$$+ \int_{\Gamma_2} wa\psi d\Gamma = \int_\Omega wQ_\psi d\Omega + \int_{\Gamma_2} w(a\psi_2 - b)d\Gamma$$

9.6.2 Spatial discretization

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

$$\Omega \approx \Omega^h \equiv \bigcup_e \Omega^e$$
(9-85)

On any finite-element domain Ω^e, the unknown vari-

able ψ (and dependent coefficients) are replaced by a *continuous approximation* that assumes the separability of space and time, thus

$$\psi(\boldsymbol{x}, t) \approx \psi^h(\boldsymbol{x}, t) = N_i(\boldsymbol{x})\psi_i(t) \tag{9-86}$$

where $i = 1, \dots, M$ designates nodal indices, M is the total number of nodes, N_i is the nodal basis function, called the trial space, and $\boldsymbol{x}$ are the spatial coordinates (9-11). Note that the summation convention is used for repeated indices. For the present analysis the basis functions N_i are based on C_0 (continuous) piece-wise polynomials that are piecewise-continuously differen-

tiable 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. (9-84) leads to the following global matrix system of M equations

$$\boldsymbol{O} \cdot \dot{\psi} + \boldsymbol{K} \cdot \psi - \boldsymbol{F} = 0 \tag{9-87}$$

with its components written in indicial notation

$$
\left.
\begin{aligned}
O_{ij} &= \sum_e O_{ij}^e = \sum_e \int_{\Omega^e} S N_i N_j d\Omega \\
K_{ij} &= \sum_e K_{ij}^e = \sum_e \left[\int_{\Omega^e} (N_i \boldsymbol{q} \cdot \nabla N_j + \nabla N_i \cdot (\boldsymbol{D} \cdot \nabla N_j) + \Phi N_i N_j) d\Omega + \int_{\Gamma_2^e} a N_i N_j d\Gamma \right] \\
F_i &= \sum_e F_i^e = \sum_e \left[\int_{\Omega^e} N_i Q d\Omega + \int_{\Gamma_2^e} N_i (a\psi_2 - b) d\Gamma \right]
\end{aligned}
\right\}
\tag{9-88}
$$

where the subscripts $i, j = 1, \dots, M$ denote nodal indices. The superposed dot in (9-87) means differentiation with respect to time t, *viz.*,

$$\dot{\psi}(t) = \left\{ \frac{d}{dt}\psi(t) \right\} \tag{9-89}$$

9.6.3 Temporal discretization

The spatially discretized equation (9-87) as a common first-order differential equation in time can practi-

cally only solved by numerical schemes. For stability reasons implicit (A-stable) two-step techniques are preferred.

Considering $\psi(t)$ within the finite interval $(t_n, t_n + \Delta t_n)$, where the subscript n denotes the time plane and Δt_n is a variable time step length, the function $\psi(t)$ is defined as

$$\psi^n = \psi(t_n) \tag{9-90}$$

at the previous (old) time plane and as

$$\psi^{n+1} = \psi(t_n + \Delta t_n) \qquad (9\text{-}91)$$

at the new time plane.

9.6.3.1 θ -Method

Introducing a weighting coefficient $(0 \le \theta \le 1)$, we can write

$$\left. \begin{array}{l} \psi(t_n + \theta\Delta t_n) = \theta\psi(t_n + \Delta t_n) + (1-\theta)\psi(t_n) \\ F(t_n + \theta\Delta t_n) = \theta F(t_n + \Delta t_n) + (1-\theta)F(t_n) \\ \dot{\psi}(t_n + \theta\Delta t_n) = \theta\dot{\psi}(t_n + \Delta t_n) + (1-\theta)\dot{\psi}(t_n) \end{array} \right\} \quad (9\text{-}92)$$

Using a backward difference approximation for $\dot{\psi}(t_n + \Delta t_n)$ and a forward difference for $\dot{\psi}(t_n)$ one obtains

$$\dot{\psi}(t_n + \theta\Delta t_n) = \frac{\psi^{n+1} - \psi^n}{\Delta t_n} \qquad (9\text{-}93)$$

Common time stepping schemes result if choosing θ in an appropriate manner, *viz.*,

$$\left. \begin{array}{lll} \theta = 0 & \text{explicit scheme} \\ \theta = 1/2 & \text{trapezoid rule (Crank-Nicolson scheme)} \\ \theta = 1 & \text{implicit scheme} \end{array} \right\} \quad (9\text{-}94)$$

Inserting (9-92) into (9-87) the following matrix equation finally results

$$\left(\frac{O}{\Delta t_n} + K\theta\right)\psi^{n+1} = \left(\frac{O}{\Delta t_n} - K(1-\theta)\right)\psi^n \qquad (9\text{-}95)$$

$$+ (F^{n+1}\theta + F^n(1-\theta))$$

9.6.3.2 Predictor-corrector method

The predictor-corrector method is thoroughly described elsewhere[3,4,5,6]. For the present analysis the fully implicit backward Euler (BE) scheme with a first-order accuracy and the semi-implicit nondissipative trapezoid rule (TR) with a second-order accuracy are enforced. The time derivatives are approximated, for the BE scheme, by

$$\dot{\psi}^{n+1} = \frac{\psi^{n+1} - \psi^n}{\Delta t_n} \qquad (9\text{-}96)$$

and for the TR scheme, by

$$\dot{\psi}^{n+1} = \frac{2}{\Delta t_n}(\psi^{n+1} - \psi^n) - \dot{\psi}^n . \qquad (9\text{-}97)$$

Inserting (9-96) and (9-97) into (9-87) results in

$$\left(\frac{O}{\theta\Delta t_n} + K\right)\psi^{n+1} = O\left[\frac{\psi^n}{\theta\Delta t_n} + \left(\frac{1}{\theta} - 1\right)\dot{\psi}^n\right] + F^{n+1} \quad (9\text{-}98)$$

with $\theta \in (\frac{1}{2}, 1)$ for the TR and BE scheme, respectively.

9.6.4   Finite-element basis operations

A fundamental aspect of the finite-element method is the use of *master elements* where all element-data inner products and integrations are performed in gener-alized (local) coordinates (see Fig. 9.6). The coordinate transformation (or mapping) that bridges a computational (transform) η-space and the Euclidean space R^D is
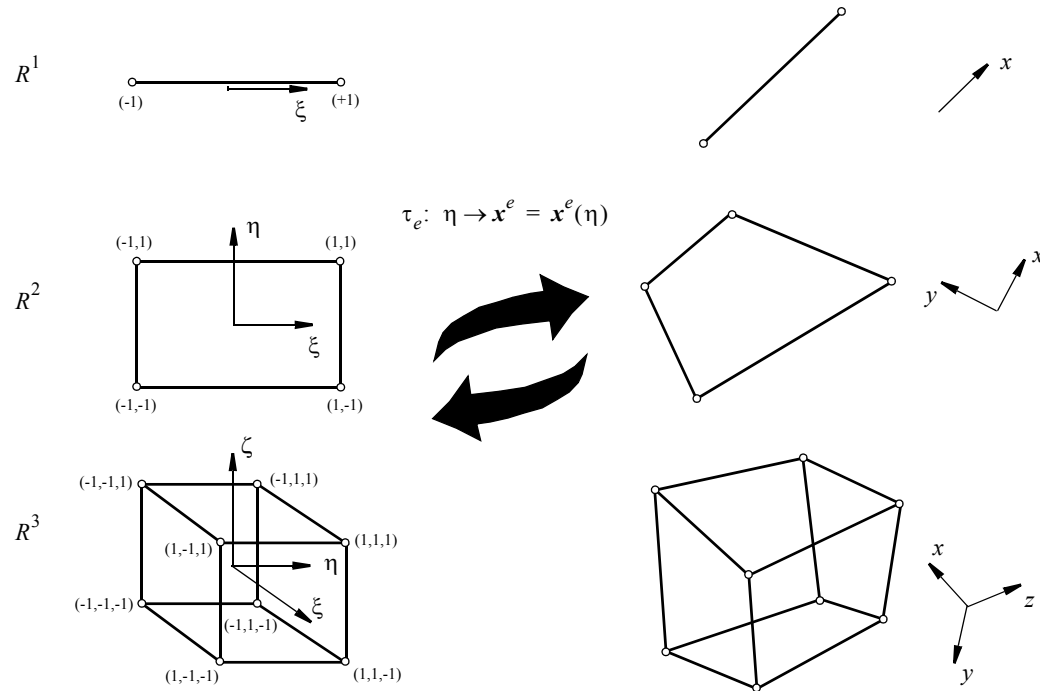

Figure 9.6 Finite elements with one-to-one mapping onto R^D ($D = 1,2,3$).

$$
\tau_e\colon\ \eta \to x = x(\eta)
$$

$$
\eta = \begin{bmatrix} \xi \\ \eta \\ \zeta \end{bmatrix} \quad \begin{array}{l} -1 \le \xi \le 1 \\ -1 \le \eta \le 1 \\ -1 \le \zeta \le 1 \end{array} \tag{9-99}
$$

Based on (9-99) it is convenient to express the basis functions N_i in local η-coordinates for each element e, *viz.*,

$$N_i = N_i(\boldsymbol{x}) = \bigcup_e N_i^e(\eta)$$

$$\boldsymbol{x} = N_i \boldsymbol{x}_i$$

(9-100)

with $\boldsymbol{x} = (x_1, x_2, x_3)$ according to (9-11). To evaluate the flux vector divergence terms in (9-88) the inverse Jacobian is required

$$\nabla N_i = \boldsymbol{J}^{-1} \left\{ \begin{array}{c} \dfrac{\partial N_i}{\partial \xi} \\[2mm] \dfrac{\partial N_i}{\partial \eta} \\[2mm] \dfrac{\partial N_i}{\partial \zeta} \end{array} \right\}$$

(9-102)

The mapping τ_e is *one-to-one and onto* its range provided the transformation Jacobian $\boldsymbol{J}$ is nonsingular, where in the R^3 space

$$\boldsymbol{J} = \frac{\partial \boldsymbol{x}}{\partial \eta} = \left\{ \begin{array}{c} \dfrac{\partial}{\partial \xi} \\[2mm] \dfrac{\partial}{\partial \eta} \\[2mm] \dfrac{\partial}{\partial \zeta} \end{array} \right\} \{x_1, x_2, x_3\} = \begin{bmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix}$$

(9-101)

$$= \begin{vmatrix} \dfrac{\partial x_1}{\partial \xi} & \dfrac{\partial x_2}{\partial \xi} & \dfrac{\partial x_3}{\partial \xi} \\[2mm] \dfrac{\partial x_1}{\partial \eta} & \dfrac{\partial x_2}{\partial \eta} & \dfrac{\partial x_3}{\partial \eta} \\[2mm] \dfrac{\partial x_1}{\partial \zeta} & \dfrac{\partial x_2}{\partial \zeta} & \dfrac{\partial x_3}{\partial \zeta} \end{vmatrix}$$

where

$$\boldsymbol{J}^{-1} = \frac{\partial \eta}{\partial \boldsymbol{x}} = \left\{ \begin{array}{ll} \dfrac{1}{\|\boldsymbol{J}\|} \begin{bmatrix} (J_{22}J_{33} - J_{32}J_{23}) & (J_{13}J_{32} - J_{12}J_{33}) & (J_{12}J_{23} - J_{13}J_{22}) \\ (J_{31}J_{23} - J_{21}J_{33}) & (J_{11}J_{33} - J_{13}J_{31}) & (J_{21}J_{13} - J_{23}J_{11}) \\ (J_{21}J_{32} - J_{31}J_{22}) & (J_{12}J_{31} - J_{32}J_{11}) & (J_{11}J_{22} - J_{12}J_{21}) \end{bmatrix} & \text{in} \quad R^3 \\[8mm] \dfrac{1}{\|\boldsymbol{J}\|} \begin{bmatrix} J_{22} & -J_{12} \\ -J_{21} & J_{11} \end{bmatrix} & \text{in} \quad R^2 \\[6mm] \dfrac{1}{\|\boldsymbol{J}\|} & \text{in} \quad R^1 \end{array} \right.$$

(9-103)

with the determinant of $\boldsymbol{J}$

$$\|J\| = \begin{cases} J_{11}(J_{22}J_{33} - J_{32}J_{23}) - J_{21}(J_{12}J_{33} - J_{13}J_{32}) + J_{31}(J_{12}J_{23} - J_{13}J_{22}) & \text{in} \quad R^3 \\ J_{11}J_{22} - J_{21}J_{12} & \text{in} \quad R^2 \\ J_{11} & \text{in} \quad R^1 \end{cases} \tag{9-104}$$

The master element matrices appearing in (9-87) and (9-88) are to be integrated over element volumes Ω^e and surfaces Γ^e. The integration in local coordinates becomes for a 'volume' element

$$d\Omega = \begin{cases} \begin{rcases} dxdydz = \|J\|d\xi d\eta d\zeta \\ dxdy = \|J\|d\xi d\eta \\ dx = \|J\|d\xi \end{rcases} \quad \text{Cartesian} \quad R^D(D = 1, 2, 3) \\ \\ rdrd\omega dz = 2\pi\|J\|rd\xi d\eta \quad \text{axisymmetric} \end{cases} \tag{9-105}$$

and for an 'areal' element in Cartesian coordinates of $R^D(D = 1, 2, 3)$ space:

$$
d\Gamma = \begin{cases}
\left\| \begin{bmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \\ \frac{\partial z}{\partial \xi} \end{bmatrix} \times \begin{bmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \\ \frac{\partial z}{\partial \eta} \end{bmatrix} \right\| d\xi d\eta = \left| \det \begin{bmatrix} i & j & k \\ J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \end{bmatrix} \right| d\xi d\eta = \left\| \begin{bmatrix} J_{12}J_{23} - J_{13}J_{22} \\ J_{13}J_{21} - J_{11}J_{23} \\ J_{11}J_{22} - J_{12}J_{21} \end{bmatrix} \right\| d\xi d\eta & \text{at} \quad \zeta = \pm 1 \\[4em]

\left\| \begin{bmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \\ \frac{\partial z}{\partial \eta} \end{bmatrix} \times \begin{bmatrix} \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \zeta} \end{bmatrix} \right\| d\eta d\zeta = \left| \det \begin{bmatrix} i & j & k \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} \right| d\eta d\zeta = \left\| \begin{bmatrix} J_{22}J_{33} - J_{23}J_{32} \\ J_{23}J_{31} - J_{21}J_{33} \\ J_{21}J_{32} - J_{22}J_{31} \end{bmatrix} \right\| d\eta d\zeta & \text{at} \quad \xi = \pm 1 \\[4em]

\left\| \begin{bmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \\ \frac{\partial z}{\partial \xi} \end{bmatrix} \times \begin{bmatrix} \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \zeta} \end{bmatrix} \right\| d\xi d\zeta = \left| \det \begin{bmatrix} i & j & k \\ J_{11} & J_{12} & J_{13} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} \right| d\xi d\zeta = \left\| \begin{bmatrix} J_{12}J_{33} - J_{13}J_{32} \\ J_{13}J_{31} - J_{11}J_{33} \\ J_{11}J_{32} - J_{12}J_{31} \end{bmatrix} \right\| d\xi d\zeta & \text{at} \quad \eta = \pm 1 \\[4em]

\left\| \begin{bmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \end{bmatrix} \right\| d\xi = \left\| \begin{bmatrix} J_{11} \\ J_{12} \end{bmatrix} \right\| d\xi = \sqrt{J_{11}^2 + J_{12}^2} \; d\xi & \text{at} \quad \eta = \pm 1 \\[3em]

\left\| \begin{bmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \end{bmatrix} \right\| d\eta = \left\| \begin{bmatrix} J_{21} \\ J_{22} \end{bmatrix} \right\| d\eta = \sqrt{J_{21}^2 + J_{22}^2} \; d\eta & \text{at} \quad \xi = \pm 1 \\[3em]

\cdots \Big|_{\xi = 1}^{\xi = -1}
\end{cases}
$$
(9-106)

and in cylindrical coordinates of R^2 (meridional) space

$$
d\Gamma = \begin{cases} \left\| \begin{bmatrix} \dfrac{\partial r}{\partial \xi} \\ \dfrac{\partial z}{\partial \xi} \end{bmatrix} \right\| r d\xi d\omega = 2\pi \left\| \begin{bmatrix} J_{11} \\ J_{12} \end{bmatrix} \right\| r d\xi = 2\pi \sqrt{J_{11}^2 + J_{12}^2} \; r d\xi & \text{at} \qquad \eta = \pm 1 \\[24pt] \left\| \begin{bmatrix} \dfrac{\partial r}{\partial \eta} \\ \dfrac{\partial z}{\partial \eta} \end{bmatrix} \right\| r d\eta d\omega = 2\pi \left\| \begin{bmatrix} J_{21} \\ J_{22} \end{bmatrix} \right\| r d\eta = 2\pi \sqrt{J_{21}^2 + J_{22}^2} \; r d\eta & \text{at} \qquad \xi = \pm 1 \end{cases}
$$

(9-107)

where $r = N_i(\eta) r_i$. Commonly, for 2D and 3D element entities the volume and area integrals are evaluated via

a Gaussian quadrature rule, e.g.,

$$
\int_\Omega f(\boldsymbol{x}) d\Omega = \sum_e \int_{\Omega_e} f(\eta) d\Omega = \sum_e \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} f^*(\xi, \eta, \zeta) d\xi d\eta d\zeta = \sum_e \sum_{i=1}^{n_{\text{Gauss}}} \sum_{j=1}^{n_{\text{Gauss}}} \sum_{k=1}^{n_{\text{Gauss}}} w_i w_j w_k f^*(\xi_i, \eta_j, \zeta_k)
$$

$$
\int_\Gamma g(\boldsymbol{x}) d\Gamma = \sum_e \int_{\Gamma_e} g(\eta) d\Gamma = \sum_e \int_{-1}^{1}\int_{-1}^{1} g^*(\xi, \eta) d\xi d\eta = \sum_e \sum_{i=1}^{n_{\text{Gauss}}} \sum_{j=1}^{n_{\text{Gauss}}} w_i w_j g^*(\xi_i, \eta_j)
$$

(9-108)

where the n_{Gauss} is the number of Gauss points, w_i are weighting coefficients and the indices (i, j, k) indicate the positions of the evaluation points in their local coordinates η. The functions $f(.)$ and $g(.)$ in the integrands are marked by an asterisk if the volume and surface integrals are expressed in the η-coordinates according to (9-105), and .

For 1D elements the integrals of (9-88) can easily evaluated in a direct analytical manner so as shown in Appendix B for a channel element with a linear basis function N.

9.6.5 Assembly of the different feature elements to the global system matrix

9.6.5.1 Needs for coordinate transformation

The global matrix equations (9-87) written in the form

$$O \cdot \dot{\psi} + K \cdot \psi - F = 0$$

$$O = \sum_e O^e$$

$$K = \sum_e K^e \qquad (9\text{-}109)$$

$$F = \sum_e F^e$$

represent the *standard discrete system* resulting from the summation (*assembly*) of the elemental (*e*) matrix contributions. The integrals of the matrices and vectors O^e, K^e and F^e for each element e are performed in the local coordinates η for the corresponding Euclidean space R^D as stated above.

Under normal conditions 1D finite elements are mapped to the R^1 space, 2D elements to the R^2 space and 3D elements to the R^3 space. In such case the mapping is strictly one-to-one, that means 3 global coordinates x, y, z are transformed to 3 local coordinates ξ, η, ζ in 3D, 2 global coordinates x, y to 2 local coordinates ξ, η in 2D and 1 global coordinate x to 1 local coordinate ξ in 1D. However, when 1D and 2D discrete feature elements are generally mapped onto a 3D global space, the number of local coordinates η will be less than the number of global coordinates x and the Jacobian J (9-101) will not be any more an invertible square matrix (e.g., for the $\xi - \eta$-system of a 2D feature element mapped into the global $x - y - z$-system the third row of J contains zeros, $J_{31} = J_{32} = J_{33} = 0$, because the ζ-coordinate does not exist in 2D elements).

There is a simple way to overcome this mapping conflicts. We take into consideration that all flow and transport processes are invariant with respect to a rotation (orthogonal transformation) of the global coordinates x. Accordingly, we can arbitrarily rotate x to the x'-coordinates by using a suitable matrix of directional cosines a as

$$x' = a \cdot x$$

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \qquad (9\text{-}110)$$

Taking an appropriate rotation of the global $x - y - z$-coordinate system in such a way that the resulting local $x' - y' - z'$-system becomes aligned to the orientation of the 2D or 1D elements in the R^3 space, there will be no more an elemental contribution to the z'-direction for 2D elements and elemental contributions to the y'- and z'-directions for 1D elements (see Fig. 9.7).

The advantages of this coordinate transformation are that the corresponding Jacobian J'

$$J' = \frac{\partial x'}{\partial \eta} \qquad (9\text{-}111)$$

becomes again an invertible square matrix and the standard metric procedure can be maintained in the assembly process for the global matrix system (9-109). To ease the computations the $x' - y' - z'$-coordinate system may, in fact, be different for every element e.

$$\tau'_e: \ \eta \rightarrow \boldsymbol{x'}^e \ = \ \boldsymbol{x'}^e(\eta)$$

2D element in the R^3 space

1D element in the R^3 space
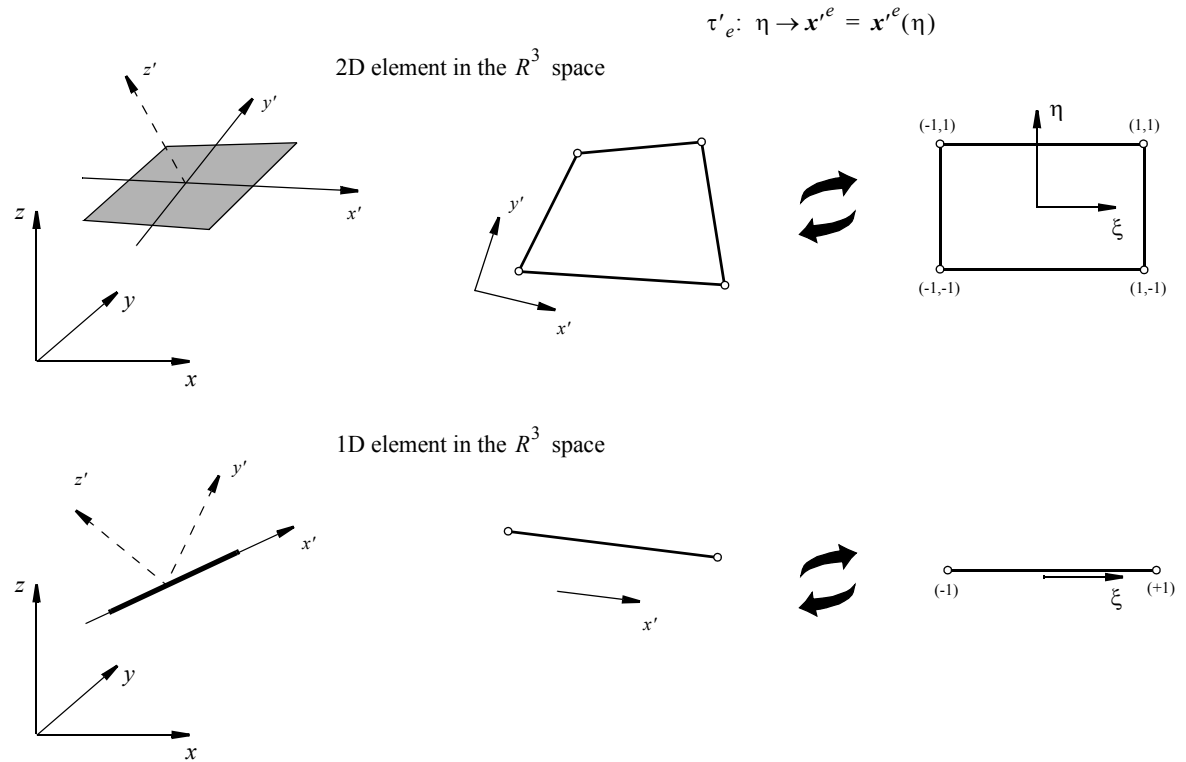
Figure 9.7 Global $x - y - z$ -coordinate system, rotated elemental $x' - y' - z'$ -coordinate system and local $\xi - (\eta)$ -coordinate system for 2D and 1D elements in the R^3 space.

9.6.5.2   Generalized discrete system

Since the rotation matrix $\boldsymbol{a}$ forms an *orthonormal basis* we can transform between the $\boldsymbol{x}$ - and the $\boldsymbol{x'}$ -system according to

$$\left.\begin{array}{ll} \boldsymbol{x'} = \boldsymbol{a}^e \cdot \boldsymbol{x} & \boldsymbol{x} = \boldsymbol{a}^{Te} \cdot \boldsymbol{x'} \\[4pt] \boldsymbol{\psi'}^e = \boldsymbol{T}^e \cdot \boldsymbol{\psi}^e & \boldsymbol{\psi}^e = \boldsymbol{T}^{Te} \cdot \boldsymbol{\psi'}^e \\[4pt] \boldsymbol{F'}^e = \boldsymbol{T}^e \cdot \boldsymbol{F}^e & \boldsymbol{F}^e = \boldsymbol{T}^{Te} \cdot \boldsymbol{F'}^e \end{array}\right\} \qquad (9\text{-}112)$$

where

$$T^e = \begin{bmatrix} a^e & 0 & 0 & \dots \\ 0 & a^e & 0 \\ 0 & 0 & a^e \\ \dots & & \end{bmatrix} \qquad (9\text{-}113)$$

is a diagonal directional cosines matrix built up of a^e matrices in a number equal to that of the nodes in the element e
and

$$T^{Te} = \begin{bmatrix} a^{Te} & 0 & 0 & \dots \\ 0 & a^{Te} & 0 \\ 0 & 0 & a^{Te} \\ \dots & & \end{bmatrix} \qquad a^{Te} = \begin{bmatrix} a^e_{11} & a^e_{21} & a^e_{31} \\ a^e_{12} & a^e_{22} & a^e_{32} \\ a^e_{13} & a^e_{23} & a^e_{33} \end{bmatrix} \quad (9\text{-}114)$$

are the *transposes* of matrices T^e and a^e, respectively.

We can usually assume that any directional properties of the discrete system (9-109) are available in the local x'-coordinates. Then, the local (elemental) system matrix $K'^e = K^e_{x'}$ is transformed into the global matrix K^e according to

$$q'^e = K'^e \cdot \psi'^e = K'^e \cdot (T^e \cdot \psi^e)$$
$$q^e = T^{Te} \cdot q'^e = [T^{Te} \cdot (K'^e \cdot T^e)] \cdot \psi^e = K^e \cdot \psi^e \qquad (9\text{-}115a)$$

with

$$K^e = T^{Te} \cdot (K'^e \cdot T^e) \qquad (9\text{-}115b)$$

where q'^e and q^e represent elemental 'flux' components for the local and the global coordinate system, respectively. On inserting (9-112), (9-115a) and (9-115b), using the same transformation rules for the storage matrix O and assuming that in general the right-hand side vector F can also be directionally dependent, the global matrix system (9-109) yields finally the form

$$\left. \begin{aligned} O \cdot \dot{\psi} + K \cdot \psi - F &= 0 \\ O &= \sum_e [T^{Te} \cdot (O'^e \cdot T^e)] \\ K &= \sum_e [T^{Te} \cdot (K'^e \cdot T^e)] \\ F &= \sum_e (T^{Te} \cdot F'^e) \end{aligned} \right\} \qquad (9\text{-}116)$$

where the rotation matrix T^e is evaluated at element level e. Practically, T^e is only required for mapping 2D and 1D feature elements in the general R^3 space, while 3D elements need not to rotate to the x'-system (the rotation matrix becomes unity $T^e \equiv I$) and can be directly mapped onto the local η-space via the Jacobian J (9-101).

However, there are important special cases to be considered here. Firstly, if the material properties of the square matrices O'^e and K'^e are independent of the coordinate directions (isotropic conditions) then we have

$$[\boldsymbol{T}^{Te} \cdot (\boldsymbol{O'}^e \cdot \boldsymbol{T}^e)] = \boldsymbol{O}^e$$

$$[\boldsymbol{T}^{Te} \cdot (\boldsymbol{K'}^e \cdot \boldsymbol{T}^e)] = \boldsymbol{K}^e$$

(9-117)

because $\boldsymbol{T}^{Te} \cdot \boldsymbol{T}^e = \boldsymbol{I}$ due to the condition of orthogonality and accordingly $\boldsymbol{O}^e = \boldsymbol{O'}^e$, $\boldsymbol{K}^e = \boldsymbol{K'}^e$.

Secondly, if the sink/source and boundary condition terms incorporated in the right-hand side vector $\boldsymbol{F}$ represent *direction-independent quantities* so as occurred in the balance equations summarized in Tables 9.5, 9.6 and 9.7, the vector $\boldsymbol{F}$ consists of nodal scalars and can be directly evaluated (no rotation), *viz.*,

$$\boldsymbol{F} = \sum_e \boldsymbol{F'}^e$$

(9-118)

9.6.5.3   Determination of the directional cosines a^e of element e

The directional cosines a^e are only required for mapping 2D and 1D discrete feature elements in the R^3 space. If we commonly assume that the 3D continuum domain Ω with its boundary $\partial\Omega$ is completely filled by 3D finite elements (e.g., hexahedral or pentahedral isoparametric elements), the 2D fracture and 1D channel elements share the nodal points of the 3D mesh and their geometric extents are aligned to surfaces, edges or diagonals of the 3D matrix elements (Fig. 9.8).

For 2D fracture elements ◁ ▢ forming surfaces of the 3D matrix element it is convenient to derive the directional cosines directly from the shape of the 3D element. We can construct the 2 directional vec-

tors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ (Fig. 9.8), which are parallel to the local ξ- and η-axes, respectively. They can be found by the following shape-derived relationships

$$\boldsymbol{u}_1 = \begin{cases} \begin{bmatrix} \dfrac{\partial x}{\partial \xi} \\[4pt] \dfrac{\partial y}{\partial \xi} \\[4pt] \dfrac{\partial z}{\partial \xi} \end{bmatrix} = \begin{bmatrix} J_{11} \\ J_{12} \\ J_{13} \end{bmatrix} \quad \text{at} \quad \zeta = \pm 1 \\[30pt] \begin{bmatrix} \dfrac{\partial x}{\partial \eta} \\[4pt] \dfrac{\partial y}{\partial \eta} \\[4pt] \dfrac{\partial z}{\partial \eta} \end{bmatrix} = \begin{bmatrix} J_{21} \\ J_{22} \\ J_{23} \end{bmatrix} \quad \text{at} \quad \xi = \pm 1 \\[30pt] \begin{bmatrix} \dfrac{\partial x}{\partial \zeta} \\[4pt] \dfrac{\partial y}{\partial \zeta} \\[4pt] \dfrac{\partial z}{\partial \zeta} \end{bmatrix} = \begin{bmatrix} J_{31} \\ J_{32} \\ J_{33} \end{bmatrix} \quad \text{at} \quad \eta = \pm 1 \end{cases}$$

(9-119)
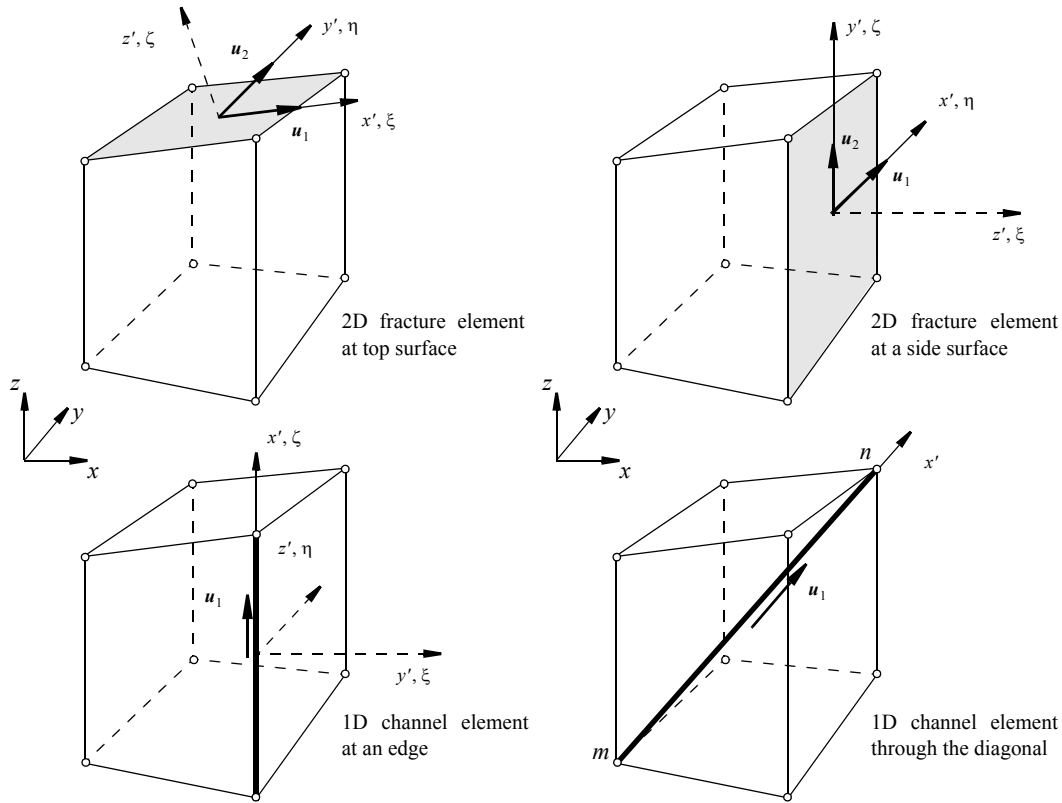
Figure 9.8 Exemplified mapping of 2D fracture elements and 1D channel elements aligned to surfaces, edges and diagonals, respectively, for a 3D finite matrix element. Local and global coordinates.

$$u_2 = \begin{cases} \begin{bmatrix} \dfrac{\partial x}{\partial \eta} \\ \dfrac{\partial y}{\partial \eta} \\ \dfrac{\partial z}{\partial \eta} \end{bmatrix} = \begin{bmatrix} J_{21} \\ J_{22} \\ J_{23} \end{bmatrix} & \text{at} \quad \zeta = \pm 1 \\[2em] \begin{bmatrix} \dfrac{\partial x}{\partial \zeta} \\ \dfrac{\partial y}{\partial \zeta} \\ \dfrac{\partial z}{\partial \zeta} \end{bmatrix} = \begin{bmatrix} J_{31} \\ J_{32} \\ J_{33} \end{bmatrix} & \text{at} \quad \xi = \pm 1 \\[2em] \begin{bmatrix} \dfrac{\partial x}{\partial \xi} \\ \dfrac{\partial y}{\partial \xi} \\ \dfrac{\partial z}{\partial \xi} \end{bmatrix} = \begin{bmatrix} J_{11} \\ J_{12} \\ J_{13} \end{bmatrix} & \text{at} \quad \eta = \pm 1 \end{cases} \quad (9\text{-}120)$$

These directional vectors can be easily used to compute the directional cosines according to

$$a_{ij}^e = \cos(u_i, e_j) = \frac{u_i^T \cdot e_j}{|u_i| \underbrace{|e_j|}_{= 1}} \quad \text{for} \quad \begin{matrix} i = 1, 2 \\ j = 1, 2, 3 \end{matrix} \quad (9\text{-}121)$$

with the basis (unit) vectors

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (9\text{-}122)$$

Notice, for 2D feature elements (fracs) we need only two directional vectors ($i = 1,2$), the remaining directional cosines a_{3j}^e are meaningless.

Often we can assume that the 2D feature elements are perfectly *flat*, i.e., they represent noncurved 2D geometries which occur for arbitrarily oriented linear triangles or for vertical linear quadrilaterals in the 3D space. Instead of using the above shape-derived expressions (9-119) and (9-120), in such cases it is convenient to derive the directional vectors u_i in a direct manner (see Fig. 9.9).

We specify the x'-axis along the edge nm of the 2D feature element. The vector u_1 is accordingly given by

$$u_1 = \begin{bmatrix} x_n - x_m \\ y_n - y_m \\ z_n - z_m \end{bmatrix} \quad (9\text{-}123)$$

The second directional vector u_2 derived by simple vector algebra (as summarized in the Appendix C) yields

$$u_2 = q - \left(\frac{q \cdot u_1}{u_1 \cdot u_1} \right) u_1 \quad (9\text{-}124a)$$

with the auxiliary vector q formed along the adjacent side lm of the 2D element as

$$q = \begin{bmatrix} x_l - x_m \\ y_l - y_m \\ z_l - z_m \end{bmatrix} \quad (9\text{-}124b)$$
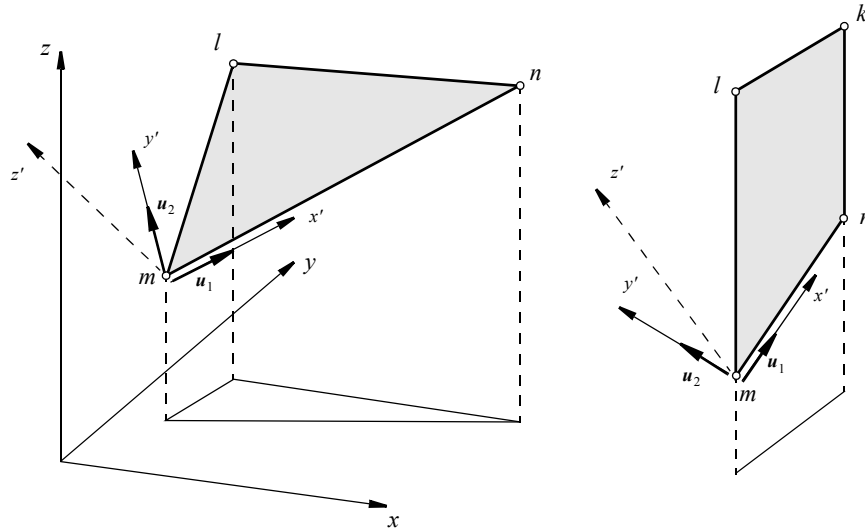
Figure 9.9 Directional vectors $\boldsymbol{u}_i$ (i = 1,2) for a linear triangular element and a vertically-oriented linear quadrilateral element.

and the directional cosines a_{ij}^e (i = 1,2; j = 1,2,3) can be easily computed by using (9-121).

For 1D channel elements the same procedure can be applied to determine a_{ij}^e (for i = 1; j = 1,2,3). Here, only one row a_{1j}^e of the rotation matrix is of interest. Taking into consideration that 1D feature elements can be rather arbitrarily placed at mesh nodes (which are not necessarily connected in one element and oriented along edges) the following direct evaluation procedure can be used to compute a_{1j}^e for a 1D linear channel element spanning between the two nodes n and m (*cf.* Fig. 9.8):

$$\left.\begin{aligned}\boldsymbol{u}_1 &= \begin{bmatrix} x_n - x_m \\ y_n - y_m \\ z_n - z_m \end{bmatrix} = \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} \\ |\boldsymbol{u}_1| &= \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \end{aligned}\right\} \quad \text{(9-125a)}$$

$$\left.\begin{aligned} a_{11}^e &= \frac{\Delta x}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \\ a_{12}^e &= \frac{\Delta y}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \\ a_{13}^e &= \frac{\Delta z}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \end{aligned}\right\} \quad \text{(9-125b)}$$

9.6.5.4   Demonstrative example

To demonstrate the assembly process for the different feature elements let us consider the simple model as shown in Fig. 9.10 consisting of only one 3D element ($e = 1$) connected with both a 2D fracture element ($e = 2$) and a 1D channel element ($e = 3$).

Node	x [unit]	y [unit]	z [unit]
1	1.	1.	0.
2	11.	1.	0.
3	7.	11.	0.
4	1.	1.	7.
5	11.	1.	9.
6	7.	11.	11.
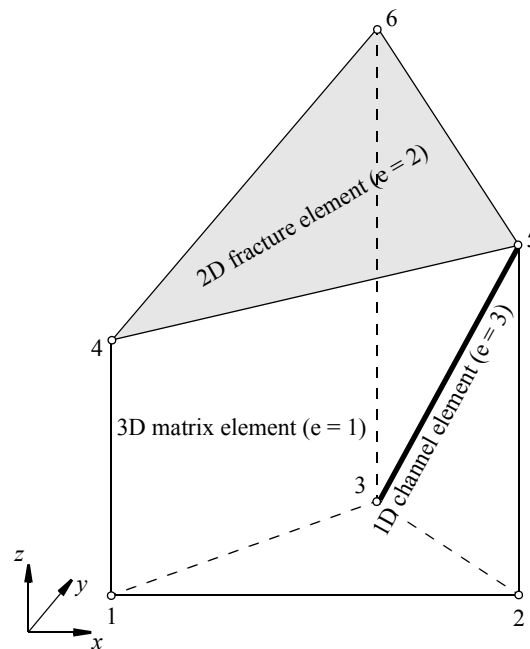
Figure 9.10 3D prismatic matrix element connected with a 2D fracture element on the top surface and a 1D channel element through a diagonal.

The assembly of the resulting matrix system has to be performed according to (9-116), *viz.*,

$$\boldsymbol{O} \cdot \dot{\psi} + \boldsymbol{K} \cdot \psi - \boldsymbol{F} = 0$$

$$\left.\begin{aligned}
\boldsymbol{O} &= [\boldsymbol{T}^{T1} \cdot (\boldsymbol{O'}^1 \cdot \boldsymbol{T}^1)] + [\boldsymbol{T}^{T2} \cdot (\boldsymbol{O'}^2 \cdot \boldsymbol{T}^2)] + [\boldsymbol{T}^{T3} \cdot (\boldsymbol{O'}^3 \cdot \boldsymbol{T}^3)] \\
\boldsymbol{K} &= [\boldsymbol{T}^{T1} \cdot (\boldsymbol{K'}^1 \cdot \boldsymbol{T}^1)] + [\boldsymbol{T}^{T2} \cdot (\boldsymbol{K'}^2 \cdot \boldsymbol{T}^2)] + [\boldsymbol{T}^{T3} \cdot (\boldsymbol{K'}^3 \cdot \boldsymbol{T}^3)] \\
\boldsymbol{F} &= (\boldsymbol{T}^{T1} \cdot \boldsymbol{F'}^1) + (\boldsymbol{T}^{T2} \cdot \boldsymbol{F'}^2) + (\boldsymbol{T}^{T3} \cdot \boldsymbol{F'}^3)
\end{aligned}\right\}$$

(9-126)

For the 3D triangular prism ($e = 1$) it is assumed that the elemental material properties can be specified in the global coordinates ($x - y - z$). Here, there is no need for coordinate transformation into the $x' - y' - z'$-coordinates. Accordingly,

$$\boldsymbol{T}^1 = \boldsymbol{T}^{T1} = \begin{bmatrix} 1 & 0 & 0 & \dots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \dots & \end{bmatrix} \qquad \boldsymbol{O}^1 = \boldsymbol{O'}^1 \qquad \boldsymbol{K}^1 = \boldsymbol{K'}^1 \qquad \boldsymbol{F}^1 = \boldsymbol{F'}^1 \qquad (9\text{-}127)$$

For the 2D triangular fracture element ($e = 2$) the directional vectors are given by

$$\boldsymbol{u}_1 = \begin{bmatrix} 10 \\ 0 \\ 2 \end{bmatrix} \qquad \boldsymbol{q} = \begin{bmatrix} 6 \\ 10 \\ 4 \end{bmatrix} \qquad \boldsymbol{u}_2 = \begin{bmatrix} 6 \\ 10 \\ 4 \end{bmatrix} - \underbrace{\left(\frac{\begin{bmatrix} 6 & 10 & 4 \end{bmatrix} \cdot \begin{bmatrix} 10 \\ 0 \\ 2 \end{bmatrix}}{\begin{bmatrix} 10 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 10 \\ 0 \\ 2 \end{bmatrix}} \right)}_{t \ = \ 0.6538} \begin{bmatrix} 10 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} -0.5385 \\ 10 \\ 2.6923 \end{bmatrix} \qquad (9\text{-}128)$$

where the formula (9-124a) is used. We note that $|\boldsymbol{u}_1| = 10.198$ and $|\boldsymbol{u}_2| = 10.370$. It leads to the following directional cosines (9-121)

$$a^2 = \begin{bmatrix} 0.9806 & 0 & 0.1961 \\ -0.0519 & 0.9843 & 0.2596 \\ . & . & . \end{bmatrix} \qquad T^2 = \begin{bmatrix} 0.9806 & 0 & 0.1961 & \dots \\ -0.0519 & 0.9843 & 0.2596 \\ . & . & . \\ \dots \end{bmatrix} \tag{9-129}$$

and we get the matrices to be assembled for the element ($e = 2$) as

$$[T^{T2} \cdot (O'^2 \cdot T^2)] = \begin{cases} \begin{bmatrix} 0.9806 & -0.0519 & . & \dots \\ 0 & 0.9843 & . \\ 0.1961 & 0.2596 & . \\ \dots \end{bmatrix} \left(\begin{bmatrix} O^2_{x'x'} & O^2_{x'y'} & 0 & \dots \\ O^2_{y'x'} & O^2_{y'y'} & 0 \\ 0 & 0 & 0 \\ \dots \end{bmatrix} \begin{bmatrix} 0.9806 & 0 & 0.1961 & \dots \\ -0.0519 & 0.9843 & 0.2596 \\ . & . & . \\ \dots \end{bmatrix} \right) & \text{if anisotropic} \\ \\ O^2 \qquad \text{if isotropic} \end{cases}$$

similar for $[T^{T2} \cdot (K'^2 \cdot T^2)]$

$$\tag{9-130}$$

$$(T^{T2} \cdot F'^2) = \begin{cases} \begin{bmatrix} 0.9806 & -0.0519 & . & \dots \\ 0 & 0.9843 & . \\ 0.1961 & 0.2596 & . \\ \dots \end{bmatrix} \begin{bmatrix} F^2_{x'} \\ F^2_{y'} \\ 0 \\ \dots \end{bmatrix} & \text{if directional} \\ \\ F^2 \qquad \text{if scalar} \end{cases}$$

Taking the u_1-vector for the 1D channel element ($e = 3$)

$$u_1 = \begin{bmatrix} 4 \\ -10 \\ 9 \end{bmatrix} \qquad |u_1| = 14.036 \tag{9-131}$$

we obtain

$$
a^3 = \begin{bmatrix} 0.2850 & -0.7125 & 0.6412 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \qquad T^3 = \begin{bmatrix} 0.2850 & -0.7125 & 0.6412 & \dots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \dots & & & \end{bmatrix}
$$

(9-132)

and finally the matrices for the element ($e = 3$) as

$$
[T^{T3} \cdot (O'^3 \cdot T^3)] = \begin{cases} \left(\begin{bmatrix} 0.2850 & \cdot & \cdot & \dots \\ -0.7125 & \cdot & \cdot \\ 0.6412 & \cdot & \cdot \\ \dots & & \end{bmatrix} \left(\begin{bmatrix} O^3_{x'x'} & 0 & 0 & \dots \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \dots & & \end{bmatrix} \begin{bmatrix} 0.2850 & -0.7125 & 0.6412 & \dots \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \dots & & \end{bmatrix} \right) \right) & \text{if anisotropic} \\ \\ O^3 & \text{if isotropic} \end{cases}
$$

similar for $[T^{T3} \cdot (K'^3 \cdot T^3)]$

(9-133)

$$
(T^{T3} \cdot F'^3) = \begin{cases} \begin{bmatrix} 0.2850 & \cdot & \cdot & \dots \\ -0.7125 & \cdot & \cdot \\ 0.6412 & \cdot & \cdot \\ \dots & & \end{bmatrix} \begin{bmatrix} F^3_{x'} \\ 0 \\ 0 \\ \dots \end{bmatrix} & \text{if directional} \\ \\ F^3 & \text{if scalar} \end{cases}
$$

References

1. Chandhry, M. H. & M. E. Barber, *Open channel flow*. In: The handbook of fluid dynamics (ed. by R. W. Johnson), CRC Press/ Springer, Chapter 45, 1998.
2. Diersch, H.-J. G., *FEFLOW - Reference Manual.* WASY Ltd., Berlin, 2002.
3. Diersch, H.-J. G., *Modeling and numerical simulation of geohydrodynamic transport processes* (in German). Reprint, WASY Ltd. Berlin, 1991.
4. Diersch, H.-J., Finite element modelling of recirculating density-driven saltwater intrusion processes in groundwater, *Advances in Water Resources* **11** (1988) 1, 25-43.
5. Diersch, H.-J. G. & O. Kolditz, Coupled groundwater flow and transport: 2. Thermohaline and 3D convection systems, *Advances in Water Resources* **21** (1998) 5, 401-425.
6. Diersch, H.-J. G. & P. Perrochet, On the primary variable switching technique for simulating unsaturated-saturated flows, *Advances in Warer Resources* **23** (1999) 3, 271-301.
7. Gottardi, G. & M. Venutelli, LANDFLOW: Computer program for the numerical simulation of two-dimensional overland flow. *Computers & Geosciences* **23** (1997)1, 77-89.
8. Gray, W.G., Derivation of vertically averaged equations describing multiphase flow in porous media. *Water Resources Research* **18** (1982)6, 1705-1712.
9. Di Giammarco, P., E. Todini & P. Lamberti, A conservative finite elements approach to overland flow: the control volume finite element formulation. Institute for Hydraulic Construction, Univ. Bologna, Italy.
10. Panton, R. L., *Incompressible flow.* 2nd Edition, J. Wiley & Sons, New York, 1996.
11. Taylor, G., Dispersion of solute matter in solvent flowing slowly through a tube. *Proc. R. Soc. London A*, **219** (1953), 186-203.

Appendix A

Nomenclature

In the above the symbols have the following meaning:

<u>Latin symbols</u>

A	=	flow area, (L^2) ;
a	=	coefficient to specify boundary conditions;
$\boldsymbol{a}$	=	rotation matrix;
B	=	thickness or depth, (L) ;
b	=	aperture or surface elevation, (L) ;
b	=	coefficient to specify boundary conditions;
C	=	concentration, (ML^{-3}) ;
C	=	Chezy roughness coefficient, $(L^{1/2}T^{-1})$;
c^f, c^s	=	specific heat capacity of fluid and solid, respectively, $(L^2 T^{-2} \Theta^{-1})$;
D	=	space dimension, (1, 2 or 3);
$\boldsymbol{D}$	=	tensor of hydrodynamic dispersion, $(L^2 T^{-1})$;
D_d	=	molecular diffusion, $(L^2 T^{-1})$;
$\boldsymbol{D}_m$	=	tensor of mechanical dispersion, $(L^2 T^{-1})$;
$\boldsymbol{d}$	=	strain-rate tensor of fluid, (T^{-1}) ;
E	=	internal (thermal) energy density, $(L^2 T^{-2})$;
$\boldsymbol{e}$	=	gravitational unit vector, (1) ;
$\boldsymbol{e}_i$	=	basis vectors, (1) ;
F	=	material surface;
f	=	specific rate of temporary production;

$f(.)$	=	general function;
f_μ	=	viscosity relation function, (1);
g	=	gravitational acceleration, (LT^{-2});
$g(.)$	=	general function;
$\boldsymbol{g}$	=	gravity vector, (LT^{-2});
h	=	hydraulic head, (L);
$\boldsymbol{I}$	=	unit (identity) tensor, (1);
$\boldsymbol{J}$	=	Jacobian matrix;
$\boldsymbol{j}$	=	flux vector;
$\boldsymbol{j}_c$	=	Fickian mass flux vector, $(ML^{-2}T^{-1})$;
$\boldsymbol{j}_T$	=	Fourierian heat flux vector, (MT^{-3});
j	=	surface or interface exchange, $(ML^{-1}T^{-2})$;
$\boldsymbol{K}$	=	$(\boldsymbol{k}\rho_o g)/\mu_o$, tensor of hydraulic conductivity, (LT^{-1});
$\boldsymbol{K}$	=	tensor function specifying different laws of flow motion, (LT^{-1});
$\boldsymbol{k}$	=	tensor of permeability, (L^2);
M	=	number of nodes;
M	=	Manning roughness coefficient, $(L^{1/3}T^{-1})$;
N	=	basis function;
$\boldsymbol{n}$	=	outward-directed normal unit vector at surface;
n_{Gauss}	=	number of Gauss points in each local coordinate direction;
p	=	fluid pressure, $(ML^{-1}T^{-2})$;
Q_c	=	mass source term, $(ML^{-3}T^{-1})$
Q_T	=	source/sink of heat, $(ML^{-1}T^{-3})$;
Q_ρ	=	fluid mass sink/source, (T^{-1});
$\boldsymbol{q}$	=	flux vector;
R	=	radius of circular tube, (L);
R^D	=	space of dimension D;

$\Re$	=	retardation, (1);
$\Re_d$	=	derivative retardation, (1);
r	=	radius, (L);
r_c	=	homogeneous chemical reaction rate, $(ML^{-3}T^{-1})$;
r_{hydr}	=	hydraulic radius, (L);
S	=	$(BS_o + S_s)$, storage term, (1);
$\boldsymbol{S}_o$	=	bed slope, inclination of the bottom plane to the horizontal x and y directions, (1);
S_o	=	compressibility, (L^{-1});
$\boldsymbol{S}_f$	=	vector of friction slopes at channel bottom, (1);
S_s	=	storativity, (1);
T, T_o	=	temperature and reference temperature, respectively, (Θ);
T_t	=	final time, (T);
t	=	time, (T);
$\boldsymbol{x}$	=	coordinate vector, (L);
x, y	=	Cartesian coordinates, (L);
z	=	axial or vertical coordinate, (L);
$\boldsymbol{v}$	=	velocity vector of fluid, (LT^{-1});
$\boldsymbol{w}$	=	velocity vector of surface or interface, (LT^{-1});
w	=	spatial weighting function;

Greek symbols

α	=	constant of friction slope relationship, (1);
β_L, β_T	=	longitudinal and transverse dispersivity, respectively, (L);
Γ_i	=	portions i of boundary $\partial\Omega$;
γ	=	fluid compressibility, (L^{-1});

9. Discrete feature modeling of flow, mass and heat transport processes by using FEFLOW

Δt_n	=	time step length at time plane n, (T);	
$\partial \Omega$	=	total boundary;	
δS	=	projected area of surface, (L^2);	
δV	=	volume of REV, (L^3);	
ε	=	porosity (= volume fraction of fluid phase), (1);	
ζ	=	$(-1 \leq \zeta \leq 1)$, local coordinate, (1);	
η	=	$(-1 \leq \eta \leq 1)$, local coordinate, (1);	
η	=	local coordinate vector, (1);	
Θ	=	$(\rho - \rho_o)/\rho_o$, density ratio or buoyancy coefficient, (1);	
θ	=	weighting coefficient, $(0 \leq \theta \leq 1)$;	
ϑ	=	chemical decay rate, (T^{-1});	
κ	=	coefficient of skeleton compressibility, (L^{-1});	
Λ	=	tensor of thermal hydrodynamic dispersion, $(MLT^{-3}\Theta^{-1})$;	
λ^f, λ^s	=	thermal conductivity for fluid and solid, respectively, $(MLT^{-3}\Theta^{-1})$;	
μ, μ_o	=	dynamic viscosity and reference dynamic viscosity of fluid, respectively, $(ML^{-1}T^{-1})$;	
ξ	=	$(-1 \leq \xi \leq 1)$, local coordinate, (1);	
ρ, ρ_o	=	fluid density and reference fluid density, respectively, (ML^{-3});	
σ	=	viscous stress tensor of fluid, $(ML^{-1}T^{-2})$;	
σ'	=	deviatory stress tensor of fluid, $(ML^{-1}T^{-2})$;	
σ^{bottom}		shear stress at bottom surface, $(ML^{-2}T^{-2})$;	
$\sigma^{interface}$		interfacial shear stress, $(ML^{-2}T^{-2})$;	
σ^{top}	=	shear stress at top surface, $(ML^{-2}T^{-2})$;	

τ	=	generalized friction factor;
Υ	=	Newton-Taylor roughness coefficient, (1);
ϕ	=	pressure head or local water depth, (L);
$\chi(C)$	=	sorption function, (1);
ψ	=	balance quantity;
Ω	=	domain;
ω	=	azimuthal angle, $(°)$;
∇	=	Nabla (vector) operator, (L^{-1});

Subscripts

corr	corrected;
e	elemental;
n	time plane;
i, j, k	nodal or spatial indices;
o	reference value;

Superscripts

α	phase index;
D	space dimension;
e	elemental;
f	fluid (water) phase;
n	time plane;
s	solid phase;
T	transpose of a matrix;
ν	surface index;

Appendix B

Analytic evaluation of matrix elements (9-88) for a 1D channel element

We consider the following linear 2-node element e
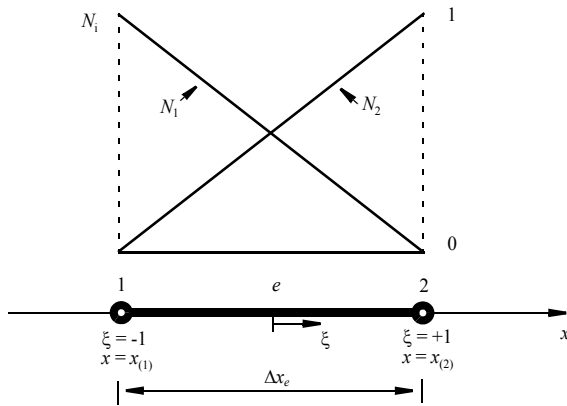
with the basis functions at the nodes 1 and 2

$$N_1 = \frac{1}{2}(1 - \xi)$$
$$N_2 = \frac{1}{2}(1 + \xi)$$
(B1)

and its derivatives

$$\frac{\partial N_1}{\partial \xi} = -\frac{1}{2}$$
$$\frac{\partial N_2}{\partial \xi} = \frac{1}{2}$$
(B2)

Furthermore, we have for the element, *cf.* (9-100),

$$x = N_1 x_{(1)} + N_2 x_{(2)}$$
(B3)

and with and (B4)

$$\|\boldsymbol{J}\| = J_{11} = \frac{\partial x}{\partial \xi} = \frac{\partial N_1}{\partial \xi} x_{(1)} + \frac{\partial N_2}{\partial \xi} x_{(2)} = \frac{\Delta x_e}{2}$$
(B4)

and

$$J^{-1} = \frac{1}{\|\boldsymbol{J}\|} = \frac{2}{\Delta x_e}$$
(B5)

Then, the divergence terms (9-102) becomes with (B2)

$$\begin{bmatrix} \nabla N_1 \\ \nabla N_2 \end{bmatrix} = J^{-1} \begin{bmatrix} \dfrac{\partial N_1}{\partial \xi} \\ \dfrac{\partial N_2}{\partial \xi} \end{bmatrix} = \begin{bmatrix} -\dfrac{1}{\Delta x_e} \\ \dfrac{1}{\Delta x_e} \end{bmatrix}$$
(B6)

According to (9-105)

$$d\Omega = dx = \|\boldsymbol{J}\| d\xi = \frac{\Delta x_e}{2} d\xi$$
(B7)

the matrices (9-88) become for element e

9. Discrete feature modeling of flow, mass and heat transport processes by using FEFLOW

$$\boldsymbol{O}^e = \int_{\Omega_e} S^e \begin{bmatrix} N_1 N_1 & N_1 N_2 \\ N_2 N_1 & N_2 N_2 \end{bmatrix} d\Omega = \frac{S^e}{4} \int_{-1}^{1} \begin{bmatrix} (1-\xi)^2 & (1-\xi^2) \\ (1-\xi^2) & (1+\xi)^2 \end{bmatrix} \frac{\Delta x_e}{2} d\xi = \frac{S^e \Delta x_e}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{B8}$$

$$\boldsymbol{K}^e = \boldsymbol{K}_1^e + \boldsymbol{K}_2^e + \boldsymbol{K}_3^e + \boldsymbol{K}_4^e$$

$$\boldsymbol{K}_1^e = \int_{\Omega_e} q^e \begin{bmatrix} N_1 \nabla N_1 & N_1 \nabla N_2 \\ N_2 \nabla N_1 & N_2 \nabla N_2 \end{bmatrix} d\Omega = \frac{q^e}{2\Delta x_e} \int_{-1}^{1} \begin{bmatrix} -(1-\xi) & (1-\xi) \\ -(1+\xi) & (1+\xi) \end{bmatrix} \frac{\Delta x_e}{2} d\xi = \frac{q^e}{2} \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$\boldsymbol{K}_2^e = \int_{\Omega_e} D^e \begin{bmatrix} \nabla N_1 \nabla N_1 & \nabla N_1 \nabla N_2 \\ \nabla N_2 \nabla N_1 & \nabla N_2 \nabla N_2 \end{bmatrix} d\Omega = \frac{D^e}{\Delta x_e^2} \int_{-1}^{1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \frac{\Delta x_e}{2} d\xi = \frac{D^e}{\Delta x_e} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{B9}$$

$$\boldsymbol{K}_3^e = \int_{\Omega_e} \Phi^e \begin{bmatrix} N_1 N_1 & N_1 N_2 \\ N_2 N_1 & N_2 N_2 \end{bmatrix} d\Omega = \frac{\Phi^e}{4} \int_{-1}^{1} \begin{bmatrix} (1-\xi)^2 & (1-\xi^2) \\ (1-\xi^2) & (1+\xi)^2 \end{bmatrix} \frac{\Delta x_e}{2} d\xi = \frac{\Phi^e \Delta x_e}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\boldsymbol{K}_4^e = a^e \begin{bmatrix} N_1 N_1 & N_1 N_2 \\ N_2 N_1 & N_2 N_2 \end{bmatrix} \Bigg|_{\xi = \xi_2 = 1}^{\xi = \xi_1 = -1} = \frac{a^e}{4} \begin{bmatrix} (1-\xi_1)^2 & (1-\xi_1^2) \\ (1-\xi_2^2) & (1+\xi_2)^2 \end{bmatrix} = a^e \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{F}^e = \boldsymbol{F}_1^e + \boldsymbol{F}_2^e$$

$$\boldsymbol{F}_1^e = \int_{\Omega_e} Q^e \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} d\Omega = \frac{Q^e}{2} \int_{-1}^{1} \begin{bmatrix} (1-\xi) \\ (1+\xi) \end{bmatrix} \frac{\Delta x_e}{2} d\xi = \frac{Q^e \Delta x_e}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{B10}$$

$$\boldsymbol{F}_2^e = (a^e \psi_2 - b^e) \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} \Bigg|_{\xi = \xi_2 = 1}^{\xi = \xi_1 = -1} = \frac{(a^e \psi_2 - b^e)}{2} \begin{bmatrix} 1 - \xi_1 \\ 1 + \xi_2 \end{bmatrix} = (a^e \psi_2 - b^e) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Finally, the discretized matrix equation (9-88) can be summarized as

$$
\sum_e \left(\frac{S^e \Delta x_e}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \cdot \left\{ \begin{matrix} \dot{\psi}_1^e \\ \dot{\psi}_2^e \end{matrix} \right\} + \left(\frac{q^e}{2} \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} + \frac{D^e}{\Delta x_e} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{\Phi^e \Delta x_e}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} + a^e \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \cdot \left\{ \begin{matrix} \psi_1^e \\ \psi_2^e \end{matrix} \right\} - \right.
$$

$$
\left. - \frac{Q^e \Delta x_e}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - (a^e \psi_2 - b^e) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) = \{0\}
$$

(B11)

Appendix C

Derivation of the orthogonal directional vectors u_i for a 2D 'flat' element

For a linear 2D 'flat' element we can find the following vectors for a typical element with the nodes $mnl(k)$:

$$
p \equiv u_1 = \begin{bmatrix} x_n - x_m \\ y_n - y_m \\ z_n - z_m \end{bmatrix} \qquad q = \begin{bmatrix} x_l - x_m \\ y_l - y_m \\ z_l - z_m \end{bmatrix}
$$

(C1)

Since the vector $u_2 \equiv q - v$ is perpendicular to the vector $p \equiv u_1$ the dot product yields

$$
(q - v) \cdot p = 0
$$

(C2)

Using the parametric description of the vector v as

$$
v = tp
$$

(C3)
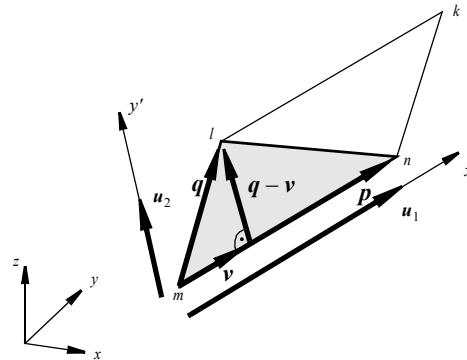
the parameter t can be easily found if inserting (C3) into (C2)

$$
\left. \begin{aligned} (q - tp) \cdot p &= 0 \\ t &= \frac{q \cdot p}{p \cdot p} \end{aligned} \right\}
$$

(C4)

As a result, we get

$$
v = \left(\frac{q \cdot p}{p \cdot p} \right) p
$$

(C5)

and finally (with $p \equiv u_1$)

$$
\left.
\begin{aligned}
u_2 &\equiv q - v = q - \left(\frac{q \cdot p}{p \cdot p}\right)p \\
u_2 &= q - \left(\frac{q \cdot u_1}{u_1 \cdot u_1}\right)u_1
\end{aligned}
\right\}
$$

(C6)

Appendix D

Input of cross-sectional data, standard implementations of hydraulic radii and their relations to different types of fractures for the Hagen-Poiseuille and the Manning-Strickler friction laws

In FEFLOW a reduced data set is used to input the geometric relationships of the fractures. There is no direct input of the hydraulic radii r_{hydr} for the different types of fracture elements applied to both the Hagen-Poiseuille and the Manning-Strickler law of fluid motion. Instead, the geometry of the cross-sections is input by the flow area A for 1D elements and by the thickness/flow depth B for 2D fracture elements (see Table 9.8).

Table 9.8 Geometric input parameter

Dimension	Fracture type/case (Table 9.5)	Parameter
1D	1DPP, 1DPN, 1DAP, 1DAN	flow area, A
2D	2DPP, 2DPN	thickness, B

On the other hand, the friction laws for the Hagen-Poiseuille flow in form of Eq. (9-56) and for the Manning-Strickler flow in form of Eq. (9-67) (with Table 9.4) are specified by the hydraulic aperture b and the friction parameter M, respectively (*cf.* Table 9.9).

Table 9.9 Frictional input parameter

Law	Parameter
Hagen-Poiseuille	hydraulic aperture, b
Manning-Strickler	roughness, M

Based on the input parameters b, A or B in FEFLOW the following relationships of the hydraulic radius r_{hydr} are implemented according to the dimension of the fracture elements. They represent so-called standard implementations as summarized in Table 9.10.

Table 9.10 Implemented standard hydraulic radii

Dimension of fracture elements	r_{hydr} expressed by appropriate input parameters	
	Hagen-Poiseuille	Manning-Strickler
1D (plane)	$b/2$ (submerged slit plane, type B of Table 9.3)	$\sqrt{A}/4$ (submerged quadratic cross-section)
2D (plane)		$B/2$ (submerged slit plane, type B of Table 9.3)

As seen from Table 9.10 the input parameter in form of the hydraulic aperture b for the Hagen-Poiseuille law and the geometric input parameter in form of the flow area A or thickness B for the Manning-Strickler law are used to express r_{hydr} internally in FEFLOW. Accordingly, FEFLOW assumes that in the case of using a Hagen-Poiseuille law a submerged slit geometry is standard (both in 1D and 2D). In case of the Manning-.Strickler law a submerged quadratic cross section $(A = B^2)$ is considered the standard geometry in 1D and a submerged slit plane geometry is standard in 2D. The question arises how is it possible to differ from the standard geometry types? For example, instead of a quadratic cross-section for a 1D fracture element, a rectangular plane or an axisymmetric geometry is to be used.

To specify hydraulic radii r_{hydr} which are different to the embodied standard geometries (Table 9.10) one can input *corrections* in the hydraulic aperture b for the Hagen-Poiseuille law and in the Manning roughness coefficient M for the Manning-Strickler law. These corrected parameters can be derived in the following manner.

The standard hydraulic conductivity $\mathbf{K}$ for the Hagen-Poiseuille law is according to Eq. (9-56) and Table 9.10:

$$\mathbf{K} = \frac{r_{\text{hydr}}^2}{a}\frac{\rho_o g}{\mu_o}\mathbf{I} = \frac{b^2}{12}\frac{\rho_o g}{\mu_o}\mathbf{I} \qquad \text{(D1)}$$

where $a = 3$ for plane geometry. Furthermore, the following standard parameters are used here for water:

$\rho_o = 10^3 \text{ kg m}^{-3}$, $\mu_o = 1.3 \text{ Pa s}$, and $g = 9.81 \text{ m}^2\text{s}^{-1}$. It results a factor of $f_o = \rho_o g/\mu_o = 7.55 \cdot 10^6 \text{ m}^{-1}\text{s}^{-1}$. A hydraulic radius which is different from the standard geometry, and parameters, which are different from the standard parameter factor f_o, can be derived from the identity

$$\frac{r_{\text{hydr}}^2}{a}f = \frac{b^2}{12}f_o \qquad \text{with} \qquad f = \frac{\rho g}{\mu} \qquad \text{(D2)}$$
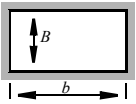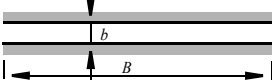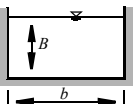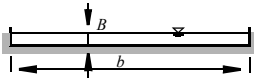
A *corrected hydraulic aperture* b_{corr} can be obtained from (D2) as

$$b_{\text{corr}} = \sqrt{\frac{f}{f_o}\frac{2\sqrt{3}}{\sqrt{a}}}r_{\text{hydr}} \qquad a = \begin{cases} 3 & \text{for plane geometry} \\ 2 & \text{for axisymmetric geometry} \end{cases} \qquad \text{(D3)}$$

where r_{hydr} is the actual (true) hydraulic radius, which can be taken from Table 9.3, and f is the true parameter factor, where dynamic viscosity, gravity and density can be specified different from the standard settings in f_o. Table 9.11 summarizes the corrected hydraulic apertures b_{corr} for interesting applications.

Table 9.11 Corrected apertures b_{corr} for different applications in the case of Hagen-Poiseuille law for 1D and 2D fracture elements [a]

	Type	b_{corr}
A	submerged rectangular cross-section 	$\dfrac{Bb}{(b+B)}\sqrt{\dfrac{f}{f_o}}$
B	submerged slit plane 	$b\sqrt{\dfrac{f}{f_o}}$ no correction is needed if $f = f_o$
C	open rectangular cross-section 	$\dfrac{2Bb}{b+2B}\sqrt{\dfrac{f}{f_o}}$
D	open wide channel ($b > 20B$) plane 	$2B\sqrt{\dfrac{f_o}{f}}$
E	submerged circular cross-section 	$\dfrac{\sqrt{3}}{\sqrt{2}}R\sqrt{\dfrac{f}{f_o}} =$ $1.224745R\sqrt{\dfrac{f}{f_o}}$

a) $f_o = \dfrac{\rho_o g}{\mu_o} = 7.55 \cdot 10^6 \text{ m}^{-1}\text{s}^{-1}$ $\qquad f = \dfrac{\rho g}{\mu}$

On the other hand, the standard hydraulic conductivity K for the Manning-Strickler law is according to Eq. (9-67) with Tables 9.4 and 9.10:

$$K = Mr_{\text{hydr}}^{2/3}\frac{I}{\sqrt[4]{\|\nabla h\|^2}} = \begin{cases} M\left(\dfrac{\sqrt{A}}{4}\right)^{2/3}\dfrac{I}{\sqrt[4]{\|\nabla h\|^2}} & \text{for 1D} \\[3mm] M\left(\dfrac{B}{2}\right)^{2/3}\dfrac{I}{\sqrt[4]{\|\nabla h\|^2}} & \text{for 2D} \end{cases} \tag{D4}$$

Accordingly, from (D4) we can find a *corrected Manning coefficient* M_{corr} in the following form to input a standard-different geometry of the fracture elements in 1D and 2D

$$M_{\text{corr}} = Mr_{\text{hydr}}^{2/3} \cdot \begin{cases} \left(\dfrac{4}{\sqrt{A}}\right)^{2/3} & \text{for 1D} \\[3mm] \left(\dfrac{2}{B}\right)^{2/3} & \text{for 2D} \end{cases} \tag{D5}$$

where M is the true (physical) Manning roughness coefficient. Tables 9.12 and 9.13 summarize the corrections M_{corr}/M for the Manning coefficients for 1D and 2D fracture elements, respectively, which are required when fracture cross-geometries different to the standard geometry have to be input.

Table 9.12 Corrected Manning roughness coefficient M_{corr} for different applications in the case of Manning-Strickler law for 1D fracture elements
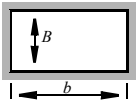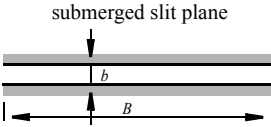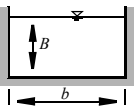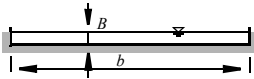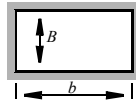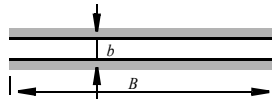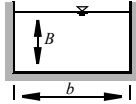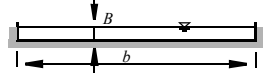
	Type	M_{corr}/M
A	submerged rectangular cross-section	$\left(\dfrac{2Bb}{(b+B)\sqrt{A}}\right)^{2/3}$ Note if $b = B$ and $A = B^2$ then $M_{corr} = M$, i.e., no correction is needed
B	submerged slit plane	$\left(\dfrac{2b}{\sqrt{A}}\right)^{2/3}$
C	open rectangular cross-section	$\left(\dfrac{4Bb}{(b+2B)\sqrt{A}}\right)^{2/3}$
D	open wide channel ($b > 20B$) plane	$\left(\dfrac{4B}{\sqrt{A}}\right)^{2/3}$
E	submerged circular cross-section	$\left(\dfrac{2R}{\sqrt{A}}\right)^{2/3}$

Table 9.13 Corrected Manning roughness coefficient M_{corr} for different applications in the case of Manning-Strickler law for 2D fracture elements

	Type	M_{corr}/M
A	submerged rectangular cross-section	$\left(\dfrac{b}{b+B}\right)^{2/3}$
B	submerged slit plane	1 no correction is needed
C	open rectangular cross-section	$\left(\dfrac{2b}{b+2B}\right)^{2/3}$
D	open wide channel ($b > 20B$) plane	$2^{2/3} = 1.5874$

9. Discrete feature modeling of flow, mass and heat transport processes by using FEFLOW

Chemical reactions

<div style="text-align:right">

10

</div>

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

10.1 Introduction

FEFLOW[4] extends the capabilities of modeling reversible and irreversible chemical reactions for single-species and multispecies systems of the flow and transport in porous media under variably saturated conditions. It encompasses the following types of reactions:

binary ion exchange reaction between fluid and solid phase (adsorption isotherms)

$$A + B \leftrightarrow C + D \qquad (10\text{-}1)$$

first-order reaction (decay)

$$A \rightarrow P \qquad (10\text{-}2)$$

consecutive reaction (decay chains, serial reaction)

$$A \rightarrow B \rightarrow C \qquad (10\text{-}3)$$
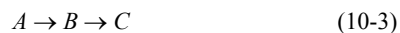
Michaelis-Menten mechanism (Monod kinetics)

$$A + E \leftrightarrow AE \rightarrow P + E \qquad (10\text{-}4)$$
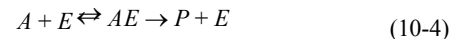
where A, B, … represent chemical species (reactants, products), the symbol $\rightarrow$ identifies the direction of reaction, $\leftrightarrow$ is used for reversible reactions at chemical equilibrium.

In the following the basics of the reactive flow and transport used in FEFLOW are summarized. It ends up with a description of the fundamental transport equations, options of chemical reactions and related rate constants available in FEFLOW.

10.2 Governing transport equations

10.2.1 Balance statements

The mass conservation of chemical species in the fluid f and solid s phases of a porous medium can be concisely written in the following form[3,4,6]

$$\varepsilon_\alpha(R^\alpha_{\text{hom}_k} + I^\alpha_k) = L^\alpha_k C^\alpha_k = \begin{cases} \dfrac{\partial}{\partial t}(\varepsilon_\alpha C^\alpha_k) - \nabla \cdot (\boldsymbol{D}^\alpha_k \cdot \nabla C^\alpha_k) + \nabla \cdot (\boldsymbol{q}^\alpha C^\alpha_k) - \varepsilon_\alpha Q^\alpha_k & \begin{aligned} \alpha &= \text{fluid} \\ k &= \text{fluid species} \end{aligned} \\[2ex] \dfrac{\partial}{\partial t}(\varepsilon_\alpha C^\alpha_k) - \varepsilon_\alpha Q^\alpha_k & \begin{aligned} \alpha &= \text{solid} \\ k &= \text{solid species} \end{aligned} \end{cases} \quad (10\text{-}5)$$

where each species, labeled by the subscript k, is associated with a particular phase $\alpha \in (f, s, \dots)$, where f and s indicate the fluid and the solid phase, respectively. In (10-5) it is

C^α_k = concentration of species k of α-phase;

$\boldsymbol{D}^\alpha_k$ = tensor of hydrodynamic dispersion;

I^α_k = diffusive flux of species k across the phase interfaces;

k = species indicator, $k = 1, \dots, N$;

L^α_k = operator as defined by Eq. (10-5);

N = total number of species;

$\boldsymbol{q}^\alpha$ = Darcy flux of α-phase;

Q^α_k = zero-order nonreactive production term of α-phase;

$R^\alpha_{\text{hom}_k}$ = homogeneous rate of reaction of species k within the α-phase;

α = phase indicator, $\alpha = f, s, \dots$;

ε_α = volume fraction of α-phase;

10.2.2 Reaction rates and multiphase aspects

The solution of the balance equations (10-5) requires knowledge of the reaction rates for kinematically controlled reactions. Different forms of the rate law can be derived. This forms depend on the type of reaction and whether the reaction is homogeneous or heterogeneous. While *homogenous reactions* take place only in one phase (f or s), the reaction of species associated with different phases (e.g., fluid species reacts with solid species) are referred to as a *heterogeneous reaction*.

In general, if a species k exists in more than one phase α, for instance the species is exchanged between fluid f and solid s phase in an adsorption process, the transport equations (10-5) have to be summed over all phases α

$$\sum_\alpha (L^\alpha_k C^\alpha_k) = \sum_\alpha \varepsilon_\alpha R^\alpha_{\text{hom}_k} + R_{\text{het}_k} \qquad (10\text{-}6)$$

where

$$R_{\text{het}_k} = \sum_\alpha \varepsilon_\alpha I^\alpha_k \qquad (10\text{-}7)$$

is the heterogeneous reaction rate of species k accumulated over all its existing phases. Let the *bulk rate of reaction* of species k in all phases be R_k as[7]

$$R_k = \sum_\alpha \varepsilon_\alpha R^\alpha_{\text{hom}_k} + R_{\text{het}_k} \qquad (10\text{-}8)$$

the mass balance equation (10-6) for species k to be solved for a multiphase system becomes

$$\sum_\alpha (L^\alpha_k C^\alpha_k) = R_k \qquad (10\text{-}9)$$

$$k = 1, ..., N \qquad \alpha = f \text{ or } s \text{ for each } k$$

If a species k exists only in one phase, say phase α, then (10-6) is reduced to

$$L^\alpha_k C^\alpha_k = \varepsilon_\alpha R^\alpha_{\text{hom}_k} + R_{\text{het}_k} \qquad (10\text{-}10)$$

The transport processes of interest refer to a fluid phase f (liquid, water) moving through a porous medium in which the void space is variably saturated by the f-phase. Conceptually, a variably saturated media consists, at least, of three phases: water f, air (or gas) a and solid s. Thus, we find the following volume fractions

$$1 = \underbrace{\varepsilon_f + \varepsilon_a}_{\varepsilon} + \varepsilon_s \qquad (10\text{-}11)$$

where ε is identified as the *porosity* (void space) of the porous media. Since the f-phase occupies only part of the void space, the *saturation* s_f of the f-phase may be used to define the relative quantity as

$$s_f = \frac{\varepsilon_f}{\varepsilon} \qquad (0 < s_f \le 1) \qquad (10\text{-}12)$$

For *saturated media* s_f becomes unity and $\varepsilon_f = \varepsilon$ with $\varepsilon_a = 0$.

For *unsaturated conditions* $s_f < 1$ it becomes apparent that only a part of the total area of the solid is exposed to ion-exchange reactions (adsorptions). Accordingly, we can subdivide the solid volume fraction ε_s into chemically active and inactive parts, viz.,

$$\varepsilon_s = \varepsilon_{s^{\text{active}}} + \varepsilon_{s^{\text{inactive}}} \qquad (10\text{-}13)$$

and the phase indicator ranges now $\alpha = f, a, s^{\text{active}}, s^{\text{inactive}}$.

Obviously, the portion of the total surface of the solid that is in contact with the f-phase depends on ε_f. It can be assumed[2] that the ratio of the solid-liquid interface to the total area of the solid is equal to the ratio of active solid volume (i.e., solid participating in the exchange reactions) to the total volume of solid, and that each of these ratios, in turn, is equal to the ratio of the liquid-occupied portion of the void space to the total void space volume, i.e., equal to s_f. Thus,

$$\frac{A_{s^{\text{active}}}}{A_{s^{\text{active}}} + A_{s^{\text{inactive}}}} = \frac{\varepsilon_{s^{\text{active}}}}{\varepsilon_s} = f(\varepsilon_f) \approx \frac{\varepsilon_f}{\varepsilon} = s_f \qquad (10\text{-}14)$$

and we obtain

$$\varepsilon_{s^{\text{active}}} = f(\varepsilon_f)\varepsilon_s \approx s_f \varepsilon_s = s_f(1 - \varepsilon) \qquad (10\text{-}15)$$

In practical modeling of flow and transport in unsaturated media, the air-phase a is commonly assumed stagnant. Focussing on dissolved chemical species k in

the f-phase for which the exchange to the air phase is negligible and the exchange to the solid phase occurs exclusively on the chemically active solid surface, the above transport equations (10-5) have to be considered only for the f- and s^{active} -phases:

$$\alpha = f, s^{\text{active}} \qquad (10\text{-}16)$$

Considering now the two-phase system of the fluid (f) and solid (s^{active}) phase, the mass transport equation (10-6) with (10-5) for a species k can be written in the form

$$L_k^f C_k^f + L_k^s C_k^s = \varepsilon_f R_{\text{hom}_k}^f + \varepsilon_{s^{\text{active}}} R_{\text{hom}_k}^s + R_{\text{het}_k}$$

$$\frac{\partial}{\partial t}(\varepsilon_f C_k^f + \varepsilon_s C_k^s) - \nabla \cdot (\boldsymbol{D}_k^f \cdot \nabla C_k^f) + \nabla \cdot (\boldsymbol{q}^f C_k^f) - Q_k = \underbrace{\varepsilon_f R_{\text{hom}_k}^f + \varepsilon_{s^{\text{active}}} R_{\text{hom}_k}^s + R_{\text{het}_k}}_{R_k}$$

$$(10\text{-}17)$$

where $Q_k = \varepsilon_f Q_k^f + \varepsilon_{s^{\text{active}}} Q_k^s$ is introduced as a nonreactive bulk production term.

For a certain number of serial and parallel reactions $r = 1, ..., N_r$ it is useful to develop the reaction rate for a species k in dependence on homogeneous and heterogeneous types of reaction. This can be expressed in the following general form[3,6]

$$R_k = \sum_\alpha \varepsilon_\alpha R_{\text{hom}_k}^\alpha + R_{\text{het}_k} \qquad (10\text{-}18)$$

$$= \sum_\alpha \varepsilon_\alpha \sum_{r=1}^{N_r^{\text{hom}}} \nu_{kr} r_r + \sum_{r=N_r^{\text{hom}}+1}^{N_r} \nu_{kr} r_r \qquad (k = 1, ..., N)$$

where

$\quad N_r \qquad$ = total number of reactions;
$\quad N_r^{\text{hom}}$ = number of homogeneous reactions

$(N_r^{\text{hom}} \leq N_r)$;
$\quad r_r \qquad$ = rate of reaction r associated with the type of reaction r ;
$\quad \nu_{kr} \qquad$ = stoichiometric number of species k and reaction r.

Note, the number of heterogeneous reactions is $N_r - N_r^{\text{hom}}$.

10.3 Basic chemical kinetics

10.3.1 Reaction stoichiometry

The basis of the chemical modeling represents the equations of reactions r which can be written in their general stoichiometric form[3]
forward reactions k_+

$$|v_{1r}|A_1 + |v_{2r}|A_2 + \ldots + |v_{N°r}|A_{N°} \xrightarrow{k^+} |v_{(N°+1)r}|B_{(N°+1)} + |v_{(N°+2)r}|B_{(N°+2)} + \ldots + |v_{Nr}|B_N \tag{10-19}$$

backward reactions k_-

$$|v_{1r}|A_1 + |v_{2r}|A_2 + \ldots + |v_{N°r}|A_{N°} \xleftarrow{k^-} |v_{(N°+1)r}|B_{(N°+1)} + |v_{(N°+2)r}|B_{(N°+2)} + \ldots + |v_{Nr}|B_N \tag{10-20}$$
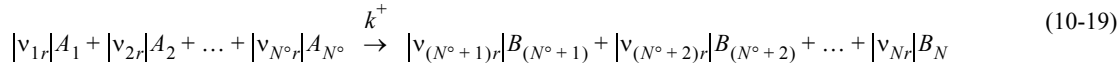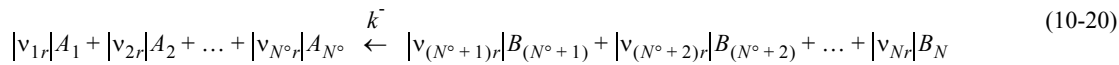
for $N° < N$ ($N°$ = number of reactants)
and $r = 1, \ldots, N_r$

which is related and quantified by the stoichiometric coefficients $|v_{kr}|$. The algebraic stoichiometric numbers v_{kr} behave:

$$v_{kr} < 0 \quad \text{for} \quad 1 \leq k \leq N° \quad \text{(reactants)}$$
$$v_{kr} > 0 \quad \text{for} \quad N° < k \leq N \quad \text{(products)}.$$

10.3.2 Rate laws and rate constants

Based on that stoichiometric form the reaction rates r_r for the forward ($r = 1$) and backward ($r = 2$) reactions can be expressed by the rate laws

$$\left.\begin{array}{l} r_1^+ = k^+ \displaystyle\prod_{k=1}^{N°} [A_k]^{|v_{k1}|} \\[2em] r_2^- = k^- \displaystyle\prod_{k=N°+1}^{N} [B_k]^{|v_{k2}|} \end{array}\right\} \tag{10-21}$$

where k^+ and k^+ are rate constants. The square bracket symbol [.] refers to the (chemical) activity of the kth species. The activity can be replaced by the product of the activity coefficient γ_k and the concentration C_k, e.g.,

$$[A_k] = \gamma_k C_k \tag{10-22}$$

The activity coefficients γ_k can be determined using empirical relationships, for example the Debye-Hückel or Davies equations[9]. For dilute solutions, γ_k approaches unity.

10.3.3 Chemical equilibrium and law of mass action (LMA)

Chemical equilibrium describes a situation in which forward and backward reactions are equal. It means

$$R_k = 0 \qquad \text{for all } k \tag{10-23}$$

or applied to the rate laws (10-21) with (10-18)

$$R_k = \begin{cases} \varepsilon_\alpha(v_{k1}r_1^+ + v_{k2}r_2^-) = 0 & \text{if homogeneous} \\ \\ v_{k1}r_1^+ + v_{k2}r_2^- = 0 & \text{if heterogeneous} \end{cases} \quad (10\text{-}24)$$

for all k

Since

$$\begin{rcases} v_{k1} = -|v_{k1}| & \text{for reactants} \\ v_{k2} = +|v_{k2}| & \text{for products} \end{rcases} \quad (10\text{-}25)$$

it is

$$|v_{k2}|r_2^- = |v_{k1}|r_1^+ \quad (10\text{-}26)$$

and with (10-21)

$$K_{eq} = \frac{|v_{k1}|k^+}{|v_{k2}|k^-} = \frac{\displaystyle\prod_{k=N°+1}^{N} [B_k]^{|v_{k2}|}}{\displaystyle\prod_{k=1}^{N°} [A_k]^{|v_{k1}|}} \quad (10\text{-}27)$$

Expression (10-27) is known as the *law of mass action* (LMA), where K_{eq} is the *equilibrium constant* to be known (measurable) for given equilibrium reactions 1 and 2.

For example, considering the more simplified equilibrium reactions in form of (10-1) in the form

$$A + B \underset{k_a^-}{\overset{k_a^+}{\rightleftarrows}} C + D \quad (10\text{-}28)$$
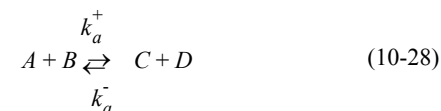
the LMA expression is

$$K_{eq} = \frac{k_a^+}{k_a^-} = \frac{[C][D]}{[A][B]} \quad (10\text{-}29)$$

10.3.4 The steady-state approximation

The steady-state approximation[1] (SSA) assumes that, during the major part of the reaction, the rates of change of concentrations of all reaction intermediates are negligibly small, i.e.,

$$R_{\text{intermediate}} \approx 0 \quad (10\text{-}30)$$

For example, considering the consecutive reaction of (10-3) in the form

$$A \overset{k_a}{\rightarrow} B \overset{k_b}{\rightarrow} C \quad (10\text{-}31)$$

the SSA applied for the species B becomes

$$k_a[A] - k_b[B] = 0 \quad (10\text{-}32)$$

10.3.5 Pre-equilibria

Considering the consecutive reactions

$$A + B \underset{k_a^-}{\overset{k_a^+}{\rightleftarrows}} C \overset{k_b}{\rightarrow} P \qquad (10\text{-}33)$$

the pre-equilibrium arises when the rates of formation of the intermediate and its decay back into reactants are much faster than its rate of formation of products. Applying the SSA scheme to species C of (10-33) it yields (exemplified for a homogeneous reaction in the α-phase)

$$R_{k=C} = \varepsilon_\alpha (k_a^+[A][B] - k_a^-[C] - k_b[C]) \approx 0 \qquad (10\text{-}34)$$

If $k_a^+ \gg k_b$ and $k_a^- \gg k_b$ it can be assumed that A, B and C are in equilibrium. Thus, the k_b-term in (10-34) vanishes and one gets

$$K_{eq} = \frac{k_a^+}{k_a^-} = \frac{[C]}{[A][B]} \qquad (10\text{-}35)$$

Then, the reaction rate for species C takes the form

$$R_{k=C} = -\varepsilon_\alpha k_b[C] = -\varepsilon_\alpha k_b K_{eq}[A][B] \qquad (10\text{-}36)$$

which represents a second-order reaction law.

On the other hand, the following reaction system

$$\left. \begin{array}{c} A + B \underset{k_a^-}{\overset{k_a^+}{\rightleftarrows}} C + D \\[2em] A \overset{k_b}{\rightarrow} P \end{array} \right\} \qquad (10\text{-}37)$$
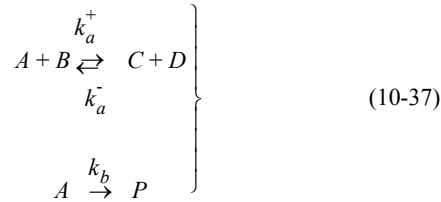
can be simplified with the pre-equilibrium assumption. The reaction rate of A simplifies to

$$R_{k=A} = \underbrace{-\varepsilon_\alpha k_a^+[A][B] + \varepsilon_\alpha k_a^-[C][D]}_{} - \varepsilon_\alpha k_b[A] \approx -\varepsilon_\alpha k_b[A] \quad (10\text{-}38)$$

$$\Rightarrow K_{eq} = \frac{k_a^+}{k_a^-} = \frac{[C][D]}{[A][B]}$$

10.4 Selected reaction processes

10.4.1 Ion-exchange reactions (adsorption isotherms)

Consider the reversible binary ion exchange (heterogeneous) reactions between the fluid species A_1^f and the sorbed (solid) species A_2^s in the form

$$|v_1|A_1^f + |v_2|A_2^s \;\leftrightarrow\; |v_1|A_1^s + |v_2|A_2^f \qquad (10\text{-}39)$$
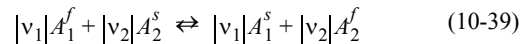
At the equilibrium the LMA (10-27) yields

10. Chemical reactions

$$K_{eq} = \frac{[A_1^s]^{|v_1|}[A_2^f]^{|v_2|}}{[A_1^f]^{|v_1|}[A_2^s]^{|v_2|}} = \frac{(\gamma_1 C_1^s)^{|v_1|}(\gamma_2 C_2^f)^{|v_2|}}{(\gamma_1 C_1^f)^{|v_1|}(\gamma_2 C_2^s)^{|v_2|}} \qquad (10\text{-}40)$$

Introducing the *ion exchange capacity* C_T^s for the sorbed species in form of

$$C_T^s = \sum_k C_k^s \qquad (10\text{-}41)$$

and the *total solution normality* for the fluid phase f as

$$C_T^f = \sum_k C_k^f \qquad (10\text{-}42)$$

Both C_T^s and C_T^f represent constant (measurable) capacities, so that for the binary ion exchange reaction (10-39) with $C_T^s = C_1^s + C_2^s$ and $C_T^f = C_1^f + C_2^f$ the LMA relationship (10-40) can be used to explicitly express the (sorbed) solid species C_1^s as a function of the (dissolved) fluid species C_1^f, *viz.*[4],

$$C_1^s = \chi_1(C_1^f) \cdot C_1^f \qquad (10\text{-}43)$$

where $\chi_1(C_1^f)$ is the adsorption function.

For monovalence $|v_1| = |v_2| = 1$ the *Langmuir*

adsorption isotherm can be directly derived[3] from (10-40). It leads to the following adsorption function

$$\chi_1(C_1^f) = \frac{k_1}{1 + k_2 C_1^f} \qquad (10\text{-}44)$$

where k_1 and k_2 are sorption coefficients. Admitting for low concentrations $k_2 C_1^f \ll 1$ the well-know *Henry adsorption isotherm* results

$$\chi_1 = k_1 \qquad (10\text{-}45)$$

Note, $k_1 = K_{eq}$ when $\dfrac{\gamma_1 C_T^s}{\gamma_2 C_T^f} \approx 1$ (see Diersch[3]).

In the case of heterovalent equilibrium reactions $|v_1| > 1, |v_2| > 1$ a polynomial expression in form of the *Feundlich adsorption isotherm* results[3]

$$\chi_1(C_1^f) = b_1(C_1^f)^{b_2 - 1} \qquad (10\text{-}46)$$

where b_1 is a sorption coefficient and b_2 is a sorption exponent.

Considering the species $k = 1$ in the fluid and active solid phase $\alpha = f, s^{active}$ the summation transport equation (10-6) and (10-17) is obtained in the form:

$$\left.\begin{aligned}
L_1^f C_1^f + L_1^s C_1^s &= \varepsilon_f R_{hom_1}^f + \varepsilon_{s^{active}} R_{hom_1}^s \\
\frac{\partial}{\partial t}[(\varepsilon_f + \varepsilon_s \chi_1)C_1^f] - \nabla \cdot (\boldsymbol{D}_1^f \cdot \nabla C_1^f) + \nabla \cdot (\boldsymbol{q}^f C_1^f) - Q_1 &= \varepsilon_f R_{hom_1}^f + \varepsilon_{s^{active}} R_{hom_1}^s
\end{aligned}\right\} \qquad (10\text{-}47)$$

which represents a single-species transport equation for the fluid species C_1^f. In it is assumed that no further

heterogeneous reactions occur, so that $R_{\text{het}_1} = 0$ according to (10-39). In a more general form it reads

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_k C_k) - \nabla \cdot (\mathbf{D} \cdot \nabla C_k) + \nabla \cdot (\mathbf{q} C_k) - Q_k \qquad (10\text{-}48)$$
$$= \varepsilon_f R^f_{\text{hom}_k} + \varepsilon_{s\,\text{active}} R^s_{\text{hom}_k}$$

in introducing the *retardation factor* $\Re_k$ (by using (10-15) and (10-12))

$$\Re_k = 1 + \frac{\varepsilon_{s\,\text{active}}}{\varepsilon_f}\chi_k = 1 + \frac{\varepsilon_s}{\varepsilon_f}f(\varepsilon_f)\chi_k \approx 1 + \frac{1-\varepsilon}{\varepsilon}\chi_k \quad (10\text{-}49)$$

with the *adsorption function* χ_k

$$\chi_k = \begin{cases} k_1 & \text{Henry} \\ \dfrac{k_1}{1 + k_2 C_k} & \text{Langmuir} \\ b_1 C_k^{b_2 - 1} & \text{Freundlich} \end{cases} \qquad (10\text{-}50)$$

where $C_k \equiv C^f_{k=1}$ is defined.

The right-hand side of and (10-48) describes irreversible homogeneous reactions of the species in the fluid and solid phase. Under exclusive reversible reactions as described by the stoichiometric equation (10-39) the homogeneous reaction rates $R^f_{\text{hom}_1}$ and $R^s_{\text{hom}_1}$ of the species in the fluid and solid phase, respectively, vanishes, i.e., $R^f_{\text{hom}_1} = R^s_{\text{hom}_1} = 0$. However, parallel to reversible reaction (10-39), reaction kinetics can occur for the species k both in the f- and the active s-phase which will be described next.

10.4.2 First-order decay reactions

Additionally to the reversible heterogeneous reaction (10-39) the species k in the f- and active s-phase should be subjected to a irreversible homogeneous first-order decay into products P according to

$$|v_k|A^f_k + |v_j|A^s_j \;\rightleftharpoons\; |v_k|A^s_k + |v_j|A^f_j \qquad (10\text{-}51a)$$

$$A^f_k \;\overset{k_a}{\rightarrow}\; P \qquad (10\text{-}51b)$$

$$A^s_k \;\overset{k_a}{\rightarrow}\; P$$
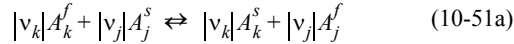
We can assume that the reversible heterogeneous reaction (10-51a) is much faster than the decay reactions (10-51b). Under such conditions the pre-equilibrium assumption (10-38) becomes applicable. It leads to the following homogeneous reaction rates for the species k in the f- and s-phase

$$\left.\begin{array}{l} R^f_{\text{hom}_k} = -k_a[A^f_k] = -k_a \gamma_k C^f_k \\[2mm] R^s_{\text{hom}_k} = -k_a[A^s_k] = -k_a \gamma_k C^s_k = -k_a \gamma_k \chi C^f_k \end{array}\right\} \qquad (10\text{-}52)$$

where the solid (sorbed) species C^s_k is replaced by the fluid species C^f_k due to the equilibrium relationship (10-43).

Inserting the reaction rates (10-52) into (10-48) the following transport equation results

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_k C_k) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C_k) + \nabla \cdot (\boldsymbol{q} C_k) - Q_k \qquad (10\text{-}53)$$

$$= -\varepsilon_f \Re_k \vartheta_k C_k$$

where the first-order decay rate ϑ_k is introduced as

$$\vartheta_k = k_a \gamma_k \qquad (10\text{-}54)$$

and the retardation factor $\Re_k$ is defined as in (10-49).

10.4.3 Michaelis-Menten mechanism

The Michaelis-Menten mechanism describes an enzyme-catalyzed reaction in which a species A is converted into products P in dependence on the concentration of the enzyme E. The mechanism is the following

$$A + E \underset{k_a^-}{\overset{k_a^+}{\longleftrightarrow}} \quad EA \quad \overset{k_b}{\rightarrow} \quad P + E \qquad (10\text{-}55)$$
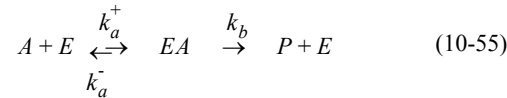
where EA denotes a bound state of the enzyme and its species. We can analyze the mechanism if assuming a pre-equilibrium for EA. The reaction rate for EA is as exemplified of a homogeneous reaction in the α-phase

$$R_{EA} = \varepsilon_\alpha \left(r_A^{k_a^+} - r_{EA}^{k_a^-} - r_{EA}^{k_b} \right) \qquad (10\text{-}56)$$

$$= \varepsilon_\alpha (k_a^+[A][E] - k_a^-[EA] - k_b[EA]) = 0$$

It follows

$$[EA] = \frac{[A][E]}{k_m} \qquad k_m = \frac{k_a^- + k_b}{k_a^+} \qquad (10\text{-}57)$$

Introducing the total concentration of enzyme as

$$[E_T] = [E] + [EA] \qquad (10\text{-}58)$$

and assuming only a little enzyme is added so that $[A]$ differs only slightly from its total concentration, then

$$[EA] = \frac{[A]([E_T] - [EA])}{k_m} \qquad (10\text{-}59)$$

which rearranges to

$$[EA] = \frac{[E_T][A]}{k_m + [A]} \qquad (10\text{-}60)$$

The reaction rate for species A in the α-phase is

$$R_A = -\varepsilon_\alpha k_a^+[A][E] + \varepsilon_\alpha k_a^-[EA] \qquad (10\text{-}61)$$

which can be simplified by applying the pre-equilibrium condition (10-56) as

$$R_A = -\varepsilon_\alpha k_b[EA] \qquad (10\text{-}62)$$

Inserting (10-60) into (10-62) it yields finally

$$R_A = -\varepsilon_\alpha \frac{k_b[E_T][A]}{k_m + [A]} = -\varepsilon_\alpha \frac{v_m C_A^\alpha}{K_m + C_A^\alpha} \qquad (10\text{-}63)$$

where

$$v_m = k_b[E_T] \quad \text{\textit{maximum velocity of enzymolysis}}$$

$$K_m = \frac{k_m}{\gamma_A} \quad \text{\textit{Michaelis (Monod) constant}}$$

(10-64)

We can generalize the reaction if we assume that the Michaelis-Menten kinetics is subjected to both the fluid and solid species k of the reversible homogeneous reaction according to (10-51a). In doing so, we are interested in the reactions

$$|v_k|A_k^f + |v_j|A_j^s \;\rightleftharpoons\; |v_k|A_k^s + |v_j|A_j^f \qquad (10\text{-}65a)$$

$$\left.\begin{array}{c} A_k^f + E \;\underset{k_a^-}{\overset{k_a^+}{\rightleftarrows}}\; EA \;\overset{k_b}{\rightarrow}\; P + E \\[2em] A_k^s + E \;\underset{k_a^-}{\overset{k_a^+}{\rightleftarrows}}\; EA \;\overset{k_b}{\rightarrow}\; P + E \end{array}\right\} \qquad (10\text{-}65b)$$
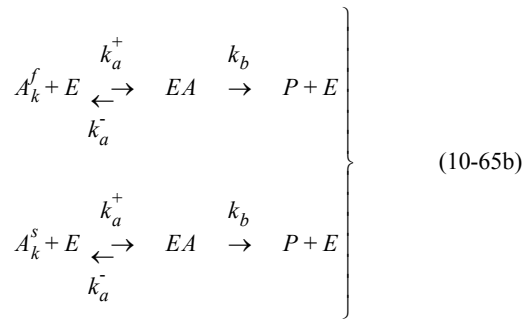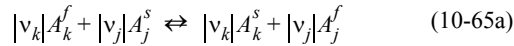
which lead to the following reaction rates

$$\left.\begin{array}{l} R_{\text{hom}_k}^f = -\dfrac{v_m C_k^f}{K_m + C_k^f} \\[2em] R_{\text{hom}_k}^s = -\dfrac{v_m C_k^s}{K_m + C_k^s} = -\dfrac{v_m \chi C_k^f}{K_m + \chi C_k^f} \end{array}\right\} \qquad (10\text{-}66)$$

Inserting the reaction rates (10-66) into (10-48) the fol-

lowing transport equation results

$$\frac{\partial}{\partial t}(\varepsilon_f \mathfrak{R}_k C_k) - \nabla \cdot (\mathbf{D} \cdot \nabla C_k) + \nabla \cdot (\mathbf{q} C_k) - Q_k \qquad (10\text{-}67)$$
$$= -\varepsilon_f (\mathfrak{R}_k + \varphi_k) \vartheta_k^m C_k$$

with the modifying function

$$\varphi_k = \frac{\chi_k - 1}{1 + \dfrac{K_m}{C_k}} \qquad (10\text{-}68a)$$

and a specific Michaelis-Menten reaction rate as

$$\vartheta_k^m = \frac{v_m}{K_m + \chi_k C_k} \qquad (10\text{-}68b)$$

where the retardation factor $\mathfrak{R}_k$ and the sorption function χ_k are according to (10-49) and (10-50), respectively.

10.4.4 Consecutive reactions

Considering consecutive reactions (termed also as decay chains or serial reactions[10], typical in radioactive decay) in the following form

$$A \overset{k_a}{\rightarrow} B \overset{k_b}{\rightarrow} C \overset{k_c}{\rightarrow} D \qquad (10\text{-}69)$$
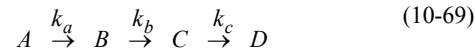
the homogeneous reaction rates for the initial reactant A, the intermediate species B and C as well as the product D in the α-phase can be written as

$$R^{\alpha}_{\text{hom}_A} = -k_a[A]$$

$$R^{\alpha}_{\text{hom}_B} = k_a[A] - k_b[B]$$

$$R^{\alpha}_{\text{hom}_C} = k_b[B] - k_c[C]$$

$$R^{\alpha}_{\text{hom}_D} = k_c[C]$$

(10-70)

Additionally, we again assume that such a type of consecutive reaction is subjected to both the fluid and solid species k of the reversible heterogeneous reaction according to (10-51a). In specifying the decay rates (10-70) in the right-hand term of the governing transport equation (10-48) we find the following set of balance equations for the four species:

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_A C_A) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C_A) + \nabla \cdot (\boldsymbol{q} C_A) - Q_A = -\varepsilon_f \Re_A \vartheta_A C_A$$

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_B C_B) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C_B) + \nabla \cdot (\boldsymbol{q} C_B) - Q_B = \varepsilon_f (\Re_A \vartheta_A C_A - \Re_B \vartheta_B C_B)$$

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_C C_C) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C_C) + \nabla \cdot (\boldsymbol{q} C_C) - Q_C = \varepsilon_f (\Re_B \vartheta_B C_B - \Re_C \vartheta_C C_C)$$

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_D C_D) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C_D) + \nabla \cdot (\boldsymbol{q} C_D) - Q_D = \varepsilon_f \Re_C \vartheta_C C_C$$

(10-71)

In a generalized form the equation system (10-71) can be written for species k of the fluid phase f as

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_k C_k) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C_k) + \nabla \cdot (\boldsymbol{q} C_k) - Q_k$$

$$= \varepsilon_f (\Re_{k-1} \vartheta_{k-1} C_{k-1} - \Re_k \vartheta_k C_k) \qquad k = 1, ..., N$$

(10-72)

Notice, in (10-72) it is by definition $\Re_0 = \vartheta_0 = C_0 \equiv 0$.

It is to be mentioned that the consecutive reactions are no more a single-species transport process. However, FEFLOW becomes capable of simulating such type of multispecies transport equations (10-72) coupled via the consecutive reaction terms by using a specific interface (IFM) programming module `chain.so`, where the equations are consecutively solved for an intermediate species $k = 2 \ (= B)$ under specific boundary conditions.

10.5 Summarized equations and relationships used in FEFLOW for modeling reactive transport processes

The reactive transport modeling is based on the following transport equation for species k dissolved in the fluid phase f

$$\frac{\partial}{\partial t}(\varepsilon_f \Re_k C_k) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C_k) + \nabla \cdot (\boldsymbol{q} C_k) - Q_k \qquad (10\text{-}73)$$

$$= \varepsilon_f [(\Re_{k-1} + \varphi_{k-1}) \vartheta_{k-1} C_{k-1} - (\Re_k + \varphi_k) \vartheta_k C_k]$$

$$(k = 1, ..., N_{\text{model}})$$

where the different reactions types and relationships are summarized in Table 10.1. The corresponding material parameters to be input for the reaction and sorption types are listed in Table 10.2.

Remarks:

(1) The *Henry isotherm* with an retardation factor of

$$\Re_k = 1 + \frac{1-\varepsilon}{\varepsilon} k_1 \qquad (10\text{-}74)$$

if often expressed by the *distribution coefficient* K_d in the form[2]

$$\Re_k = 1 + \frac{1-\varepsilon}{\varepsilon} \rho_s K_d \qquad \text{with} \qquad K_d = \frac{k_1}{\rho_s} \qquad (10\text{-}75)$$

where ρ_s is the density of solid. An alternative definition of the distribution coefficient can sometimes be found as

$$\Re_k = 1 + \frac{\rho_b K_d}{\varepsilon} \qquad (10\text{-}76)$$

where $\rho_b = (1-\varepsilon)\rho_s$ is the bulk density of the porous media (mass dry media per total volume).

(2) Microbiological degradation processes represents biologically catalyzed reactions[8] which can be appropriately modeled by *Monod kinetics*. A typical application is the oxidation of organic matter by aerobic bacteria. As long as the dissolved oxygen remains higher than a critical level the oxidation is unbounded. Below a critical level, the rate of aerobic respiration decreases with decreasing O_2 concentration, and vanishes when no O_2 is left. Thus, the limitation of aerobic respiration by O_2 obeys,

$$R_{O_2} = -v_m \frac{[O_2]}{K_{O_2} + [O_2]} \qquad (10\text{-}77)$$

where v_m is the maximum rate and $[O_2]$ is the dissolved oxygen activity. According to (10-77), when $[O_2] \ll K_{O_2}$ the availability of oxygen is limiting and the rate of aerobic respiration exhibits a first-order dependence on $[O_2]$. When, on the other hand, $[O_2] \gg K_{O_2}$, the aerobic bacteria are functioning at maximum capacity, that is, they cannot produce more enzymes to utilize the excess O_2.

Monod reaction laws are equivalent to the *Michaelis-Menten mechanism* (10-65b) (10-66) with the reaction rates summed for the fluid and solid phases

$$-\varepsilon_f (\Re_k + \varphi_k) \vartheta_k^m C_k = -\left(\frac{\varepsilon_f}{K_m + C_k} + \frac{\varepsilon_s^{\text{active}} \chi}{K_m + \chi C_k} \right) v_m C_k \quad (10\text{-}78)$$

It is a hyperbolic function that simplifies to a first-order reaction in C_k when $C_k \ll K_m$, i.e.,

$$-\varepsilon_f(\mathfrak{R}_k + \varphi_k)\vartheta_k^m C_k \;\Rightarrow\; -\varepsilon_f \mathfrak{R}_k \vartheta_k C_k \qquad (10\text{-}79)$$

$$\text{with} \qquad \vartheta_k = \frac{v_m}{K_m}$$

Such a simplistic first-order reaction ignore the fact that active microorganisms must be present to catalyze the reactions.

Table 10.1 Reaction types and relationships

Reaction type (reversible and irreversible reactions) $k = 1, \ldots, N_{model}$	N_{model}	Retardation/sorption $\mathfrak{R}_k$	Decay rate ϑ_k	Modifying function φ_k
exchange reaction with first-order decay $|v_k|A_k^f + |v_j|A_j^s \underset{\longrightarrow}{\overset{\longleftarrow}{}} |v_k|A_k^s + |v_j|A_j^f$ $A_k^f \to P$ $A_k^s \to P$	1	$\mathfrak{R}_k = 1 + \dfrac{1-\varepsilon}{\varepsilon}\chi_k$ with $\chi_k = \begin{cases} k_1 & Henry \\ \dfrac{k_1}{1+k_2 C_k} & Langmuir \\ b_1 C_k^{b_2-1} & Freundlich \end{cases}$ and $k_1 = const$ $k_2 = const$ $b_1 = const$ $b_2 = const$	$\vartheta_k = const$	$\varphi_k = 0$
exchange reaction with Michaelis-Menten kinetics $|v_k|A_k^f + |v_j|A_j^s \underset{\longrightarrow}{\overset{\longleftarrow}{}} |v_k|A_k^s + |v_j|A_j^f$ $A_k + E \overset{\longrightarrow}{\longleftarrow} EA \to P + E$ $A_k + E \overset{\longrightarrow}{\longleftarrow} EA \to P + E$	1		$\vartheta_k = \dfrac{v_m}{K_m + \chi_k C_k}$ with $v_m = const$ $K_m = const$	$\varphi_k = \dfrac{\chi_k - 1}{1 + \dfrac{K_m}{C_k}}$
exchange reaction with consecutive reaction (decay chain) $|v_k|A_k^f + |v_j|A_j^s \underset{\longrightarrow}{\overset{\longleftarrow}{}} |v_k|A_k^s + |v_j|A_j^f$ $A_{k-1}^f \to A_k^f \to P$ $A_{k-1}^s \to A_k^s \to P$	2		$\vartheta_{k-1} = const$ $\vartheta_k = const$	$\varphi_{k-1} = 0$ $\varphi_k = 0$

Table 10.2 Model input parameter for reactive transport

	Retardation/sorption				Decay				
	Henry	Langmuir		Freundlich		first-order	Michaelis-Menten		Decay chain	
Symbol	k_1	k_1	k_2	b_1	b_2	ϑ_k	v_m	K_m	ϑ_{k-1}	ϑ_k
Unit	[1]	[1]	$\left[\frac{\text{mg}}{\text{l}}\right]^{-1}$	$\left[\frac{\text{mg}}{\text{l}}\right]^{1-b_2}$	[1]	$[\text{d}^{-1}]$	$\left[\frac{\text{mg}}{\text{l}}\text{d}^{-1}\right]$	$\left[\frac{\text{mg}}{\text{l}}\right]$	$[\text{d}^{-1}]$	$[\text{d}^{-1}]$

(3) *Nonequilibrium sorption processes* can be of importance for in situ bioremediation of contaminated aquifers[5]. Typically, if groundwater extraction is applied the contaminant concentrations show a rapid initial decrease followed by a period of a much slower decrease in concentration. One of the reason is that the rate of desorption of contaminants from the aquifer solids to the fluid phase is slow, especially at low concentrations. While soil-sorbed contaminants have been shown to be generally unavailable for bioremediation, only the contaminant concentration of the fluid phase can be directly influenced in the remediation process. When desorption of the contaminant from the solid phase to the fluid phase becomes slow, the performance of in situ bioremediation is controlled by the desorption mechanism. This is called *rate-limiting desorption*, which can be described by the following nonequilibrium reaction of two different species k and j

$$A_k^f \underset{\text{desorption } k_a^-}{\overset{\text{adsorption } k_a^+}{\rightleftarrows}} A_j^s \qquad (10\text{-}80)$$
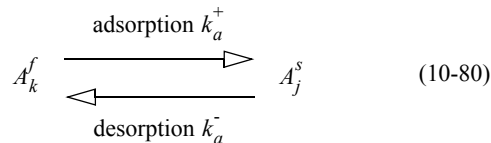
The heterogeneous reaction rates for the two species j and k in the solid and fluid phases, respectively, are

$$R_{\text{het}_j} = k_a^+[A_k^f] - k_a^-[A_j^s]$$
$$R_{\text{het}_k} = -k_a^+[A_k^f] + k_a^-[A_j^s] = -R_{\text{het}_j} \qquad (10\text{-}81)$$

Inserting the quotient of the rate constants as the equilibrium constant, cf. Eq. (10-29)

$$\frac{k_a^+}{k_a^-} = K_{\text{eq}} \qquad (10\text{-}82)$$

we get

$$R_{\text{het}_j} = k_a^-(K_{\text{eq}}[A_k^f] - [A_j^s]) = \vartheta^{\text{desorp}}(K_{\text{eq}}C_k^f - C_j^s)$$
$$R_{\text{het}_k} = -R_{\text{het}_j} \qquad (10\text{-}83)$$

where $\vartheta^{\text{desorp}} = k_a^-\gamma_j$ at $\gamma_j \approx \gamma_k$ represents now a first-order desorption rate coefficient.

The mass transport equations (10-10) for solid and fluid species are then

$$L_j^s C_j^s \;=\; \varepsilon_s R_{\mathrm{hom}_j}^s + R_{\mathrm{het}_j} \;=\; \varepsilon_s R_{\mathrm{hom}_j}^s + \vartheta^{\mathrm{desorp}}(K_{\mathrm{eq}} C_k^f - C_j^s)$$
$$L_k^f C_k^f \;=\; \varepsilon_f R_{\mathrm{hom}_k}^f - R_{\mathrm{het}_j} \;=\; \varepsilon_f R_{\mathrm{hom}_k}^f + \varepsilon_s R_{\mathrm{hom}_j}^s - L_j^s C_j^s \tag{10-84}$$

For sake of simplicity, denoting with $C = C_k^f$ the fluid concentration and with $S = C_j^s$ the sorbed concentration Eq. (10-84) with (10-5) gives the following set of mass transport equations describing the nonequilibrium sorption processes

$$\left.\begin{array}{r}\dfrac{\partial}{\partial t}(\varepsilon_s S) - \varepsilon_s Q_S^s \;=\; \varepsilon_s R_{\mathrm{hom}_S}^s + \vartheta^{\mathrm{desorp}}(K_{\mathrm{eq}} C - S) \\[2ex] \dfrac{\partial}{\partial t}(\varepsilon_f C + \varepsilon_s S) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C) + \nabla \cdot (\boldsymbol{q} C) - \varepsilon_f Q_C^f - \varepsilon_s Q_S^s \;=\; \varepsilon_f R_{\mathrm{hom}_C}^f + \varepsilon_s R_{\mathrm{hom}_S}^s\end{array}\right\} \tag{10-85}$$

which have to be solved for S and C.

References

1. Atkins, P.W., *Physical chemistry.* 5th edition. Oxford University Press, Oxford, 1994.
2. Bear, J. and Verruijt, A., *Modeling groundwater flow and pollution.* D. Reidel Publ., Dordrecht, 1987.
3. Diersch, H.-J.G., *Modeling and numerical simulation of geohydrodynamic transport processes* (in German). Reprint, WASY Ltd. Berlin, 1991.
4. Diersch, H.-J.G. Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.
5. Fry, V.A. and Istok, J.D., Effects of rate-limited desorption on the feasibility of in situ bioremediation. *Water Res. Resour.* **30** (1994)8, 2413-2422.
6. Lichtner, P.C., Continuum formulation of multicomponent-multiphase reactive transport. In: *Reactive transport in porous media*, ed. by Lichtner *et al.*, Reviews in Mineralogy, Vol. 34, Mineralogical Soc. Am., Washington DC, 1996, pp. 1-81.
7. Nguyen, V.V., Gray, W.G., Pinder, G.F., Botha, J.F. and Crerar, D.A., A theoretical investigation on the transport of chemicals in reactive porous media. *Water Res. Resour.* **18** (1982)4, 1149-1156.
8. Rittmann, B.E. and VanBriesen, J.M., Microbiological processes in reactive modeling. In: *Reactive transport in porous media*, ed. by Lichtner *et al.*, Reviews in Mineralogy, Vol. 34, Mineralogical Soc. Am., Washington DC, 1996, pp. 311-334.
9. Stumm, W. and Morgan, J.J., *Aquatic chemistry.* Wiley-Interscience, New York, 1981.
10. Sun, Y., Petersen, J.N., Clement, T.P. and Skeen, R.S., Development of analytical solutions for multispecies transport with serial and parallel reactions. *Water Res. Resour.* **35** (1999)1, 185-190.

Remarks on gas flow modeling by using FEFLOW 11

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

11.1 Basic Flow Equations

Starting point for modeling gas flows in porous media represents the following fundamental balance laws[2]:

Mass conservation (continuity equation)

$$\frac{\partial}{\partial t}(\varepsilon\rho) + \frac{\partial}{\partial x_i}(\varepsilon\rho v_i) = \varepsilon\rho Q \qquad (11\text{-}1)$$

Momentum conservation (Darcy equation)

$$\varepsilon v_i = -\frac{k_{ij}}{\mu}\left(\frac{\partial p}{\partial x_j} - \rho g_j\right) \qquad (11\text{-}2)$$

where

ε	=	porosity (= volume fraction of gas phase);
ρ	=	gas density;
v_i	=	vector of pore velocity;
Q	=	sink/source;
k_{ij}	=	permeability tensor;
μ	=	dynamic viscosity of gas;
p	=	gas pressure;
g_j	=	gravitational vector;
t	=	time;
x_i	=	Cartesian coordinates (Eulerian spatial coordinate vector);

Let us consider a *homogeneous compressible gas* the Darcy equation (11-2) for the gas motion can be simplified to

$$\varepsilon v_i = -\frac{k_{ij}}{\mu}\frac{\partial p}{\partial x_j} \qquad (11\text{-}3)$$

Introducing (11-3) into (11-1) it yields the following general model equation for the flow of a homogeneous compressible gas in a porous medium

$$\varepsilon\frac{\partial\rho}{\partial t} + \rho\frac{\partial\varepsilon}{\partial t} - \frac{\partial}{\partial x_i}\left(\frac{k_{ij}\rho}{\mu}\frac{\partial p}{\partial x_j}\right) = \varepsilon\rho Q \qquad (11\text{-}4)$$

The first two terms of the left-hand side of (11-4) can be expanded with respect to the pressure p as

11. Remarks on gas flow modeling by using FEFLOW

$$\left[\underbrace{\varepsilon\left(\frac{1}{\rho}\frac{\partial\rho}{\partial p}\right)}_{\beta_p}\rho + \rho\underbrace{\left(\frac{1}{\varepsilon_s}\frac{\partial\varepsilon}{\partial p}\right)}_{\alpha_p}\varepsilon_s\right]\frac{\partial p}{\partial t} - \frac{\partial}{\partial x_i}\left(\frac{k_{ij}\rho}{\mu}\frac{\partial p}{\partial x_j}\right) = \varepsilon\rho Q \quad (11\text{-}5)$$

or

$$\rho[\varepsilon\beta_p + (1-\varepsilon)\alpha_p]\frac{\partial p}{\partial t} - \frac{\partial}{\partial x_i}\left(\frac{k_{ij}\rho}{\mu}\frac{\partial p}{\partial x_j}\right) = \varepsilon\rho Q \quad (11\text{-}6)$$

where

β_p = gas compressibility;
α_p = skeleton (matrix) compressibility;
ε_s = $1-\varepsilon$, volume fraction of solid;

We introduce the *specific storativity* S_{0p} as

$$S_{0p} = \varepsilon\beta_p + (1-\varepsilon)\alpha_p \quad (11\text{-}7)$$

which is related to pressure changes. Finally the following model equation holds

$$\rho S_{0p}\frac{\partial p}{\partial t} - \frac{\partial}{\partial x_i}\left(\frac{k_{ij}\rho}{\mu}\frac{\partial p}{\partial x_j}\right) = \varepsilon\rho Q \quad (11\text{-}8)$$

written for the gas pressure as primary variable to be determined.

Alternatively, if introducing the piezometric head (potential) h as[2]

$$h = \frac{p}{\rho_0 g} + z \quad (11\text{-}9)$$

where

ρ_0 = reference density of gas;
g = gravitational acceleration;
z = elevation above a reference datum.

Since $\partial p/\partial t = \rho_0 g(\partial h/\partial t)$ and $\partial p/\partial x_j = \rho_0 g(\partial h/\partial x_j)$ the equation (11-8) can be expressed by the piezometric head variable h and leads to

$$\rho\rho_0 g S_{0p}\frac{\partial h}{\partial t} - \frac{\partial}{\partial x_i}\left(K_{ij}\rho\frac{\mu_0}{\mu}\frac{\partial h}{\partial x_j}\right) = \varepsilon\rho Q \quad (11\text{-}10)$$

where

K_{ij} = $\dfrac{k_{ij}\rho_0 g}{\mu_0}$, tensor of 'hydraulic' conductivity;

μ_0 = reference viscosity.

Preferring the variable h instead of the pressure p it is no more convenient to use the pressure-related specific storativity S_{0p}. Instead the compressibility coefficients can be directly derived by h according to

$$\left.\begin{aligned}\beta_h &= \frac{1}{\rho}\frac{\partial\rho}{\partial h} \\[2mm] \alpha_h &= \frac{1}{\varepsilon_s}\frac{\partial\varepsilon}{\partial h}\end{aligned}\right\} \quad (11\text{-}11)$$

which are the standard compressibility coefficients used in FEFLOW[2]. It becomes clear that the p-based and the h-based compressibilities are related as (*cf.* also Bear and Verruijt[1], p. 59)

$$\beta_h = \rho_0 g \beta_p$$
$$\alpha_h = \rho_0 g \alpha_p \tag{11-12}$$

Introducing the h-related specific storativity S_{0h}

$$S_{0h} = \varepsilon\beta_h + (1-\varepsilon)\alpha_h \tag{11-13}$$

where we note that

$$S_{0h} = \rho_0 g S_{0p} \tag{11-14}$$

we find from (11-10) the following model equation

$$\rho S_{0h}\frac{\partial h}{\partial t} - \frac{\partial}{\partial x_i}\left(K_{ij}\rho\frac{\mu_0}{\mu}\frac{\partial h}{\partial x_j}\right) = \rho Q_\rho \tag{11-15}$$

to be solved for the piezometric head h of the gas. In (11-15) we introduce the bulk sink/source rate Q_ρ defined as

$$Q_\rho = \varepsilon Q \tag{11-16}$$

which is the sink/source definition[2] commonly used in FEFLOW.

11.2 Simplifications

The general equations (11-8) and (11-15) for the pressure p and the piezometric head h of a gas flow in the porous media can be simplified for the following assumptions:

• isothermal gas flow

• constant viscosity $\mu = \mu_0$, and
• presence of a thermodynamically ideal gas.

The assumption of an ideal gas holds the following relationship for the gas density ρ:

$$\rho = \rho_0\frac{p}{p_0} \tag{11-17}$$

where

$$\rho_0 = \text{reference density of gas;}$$
$$p_0 = \text{reference pressure of gas.}$$

Inserting (11-17) into (11-8) it yields

$$p S_{0p}\frac{\partial p}{\partial t} - \frac{\partial}{\partial x_i}\left(\frac{k_{ij}p}{\mu}\frac{\partial p}{\partial x_j}\right) = p Q_\rho \tag{11-18}$$

Using (11-17) it can be easily shown that

$$\beta_p = \frac{1}{\rho}\frac{\partial\rho}{\partial p} = \frac{1}{p} \tag{11-19}$$

It can be further assumed for a gas flow that the skeleton compressibility α_p is much smaller against the gas compressibility β_p, $viz.$,

• $\alpha_p \ll \beta_p$

Accordingly, we simplify the storativity term by using (11-19) to

$$S_{0p} = \varepsilon\beta_p + (1-\varepsilon)\alpha_p \approx \varepsilon\beta_p = \frac{\varepsilon}{p} \tag{11-20}$$

and equation (11-18) yields

$$\varepsilon\frac{\partial p}{\partial t} - \frac{\partial}{\partial x_i}\left(\frac{k_{ij}p}{\mu}\frac{\partial p}{\partial x_j}\right) = pQ_\rho \tag{11-21}$$

which is nonlinear.

An alternative formulation to (11-21) is often used by introducing the square pressure p^2. Since $p\frac{\partial p}{\partial x_j} = \frac{1}{2}\frac{\partial p^2}{\partial x_j}, p\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial p^2}{\partial t}$ we obtain from (11-21)

$$\varepsilon\frac{\partial p^2}{\partial t} - p\frac{\partial}{\partial x_i}\left(\frac{k_{ij}\partial p^2}{\mu\ \partial x_j}\right) = 2p^2Q_\rho \tag{11-22}$$

which have to be solved for the square pressure p^2 as a new variable. Consider, however, equation (11-22) is still nonlinear for transient and sink/source conditions. Only for a steady-state gas flow and without the sink/source term Q_ρ a linear relationship results[4]

$$\frac{\partial}{\partial x_i}\left(\frac{k_{ij}\partial p^2}{\mu\ \partial x_j}\right) = 0 \tag{11-23}$$

Analogously, the h-based formulations can be easily derived. On this occasion the piezometric head h can be simplified for a gas flow as

$$h = \frac{p}{\rho_0 g} + z \quad \approx \quad \frac{p}{\rho_0 g} \tag{11-24}$$

because the density ρ_0 of the gas is significantly smaller than water. Using (11-24) and (11-17) with $\mu = \mu_0$, $S_{0h} \approx \varepsilon/h$ we obtain from (11-18)

$$\varepsilon\frac{\partial h}{\partial t} - \frac{\partial}{\partial x_i}\left(K_{ij}h\frac{\partial h}{\partial x_j}\right) = hQ_\rho \tag{11-25}$$

Otherwise, introducing the square potential h^2 the alternative formulation of (11-25) is

$$\varepsilon\frac{\partial h^2}{\partial t} - h\frac{\partial}{\partial x_i}\left(K_{ij}\frac{\partial h^2}{\partial x_j}\right) = 2h^2Q_\rho \tag{11-26}$$

which becomes only linear for both steady-state and divergence-free (without of sink/sources) conditions as the Laplacian of h^2, *viz.*,

$$\frac{\partial}{\partial x_i}\left(K_{ij}\frac{\partial h^2}{\partial x_j}\right) = 0 \tag{11-27}$$

11.3 Using FEFLOW for Solving the Nonlinear and Linearized Gas Flow Equations

FEFLOW[2] solves balance equations for water flow which can also be applicable directly to gas flows. Since FEFLOW prefers the piezometric head variable h as the primary unknown of the flow we find a direct analogy of the nonlinear gas flow equation (11-25) to the depth-integrated flow equations of unconfined aquifers (*cf.* Eqs. (171) to (175) in[2]). Accordingly, for 2D problems the nonlinear gas model (11-25) can be directly solved by FEFLOW. On the other hand, under steady state and without sink/sources equation (11-27) is linear in h^2 and FEFLOW can immediately applied.

Restrictions exist for the general case in solving (11-25) (or (11-26)). However, it has been shown[3] a linearization of the gas flow equation is suited for a wide range of practical applications. In the context of FEFLOW, we consider a linearized version of the governing compressible gas equation (11-25) in the form

$$\varepsilon \frac{\partial h}{\partial t} - \frac{\partial}{\partial x_i}\left(T_{ij}\frac{\partial h}{\partial x_j}\right) = hQ_\rho \qquad (11\text{-}28)$$

where the product $K_{ij}h$ is approximated by the transmissivity T_{ij}, which can be linearized by the reference potential h_0, *viz.*,

$$T_{ij} = K_{ij}h \approx K_{ij}h_0 . \qquad (11\text{-}29)$$

It is often sufficient to approximate the sink/source expression of (11-28) by

$$\overline{Q} = hQ_\rho \approx h_0 Q_\rho \qquad (11\text{-}30)$$

and we finally solve the linearized h-based gas flow equation in the form

$$\varepsilon \frac{\partial h}{\partial t} - \frac{\partial}{\partial x_i}\left(T_{ij}\frac{\partial h}{\partial x_j}\right) = \overline{Q} \qquad (11\text{-}31)$$

Alternatively, instead of linearizing (11-25) we can also apply (11-26) to solve the gas flow equation by the h^2 variable. In doing so, we obtain from (11-26) the following linearized h^2-based gas flow equation, *viz.*,

$$\overline{S}\frac{\partial h^2}{\partial t} - \frac{\partial}{\partial x_i}\left(K_{ij}\frac{\partial h^2}{\partial x_j}\right) = 2\overline{Q} \qquad (11\text{-}32)$$

where a pseudo-storativity is introduced as

$$\overline{S} \approx \frac{\varepsilon}{h_0} \qquad (h_0 \neq 0) \qquad (11\text{-}33)$$

For steady-state flow conditions we can solve either the h-based form

$$-\frac{\partial}{\partial x_i}\left(T_{ij}\frac{\partial h}{\partial x_j}\right) = \overline{Q} \qquad (11\text{-}34)$$

or the h^2-based form

$$-\frac{\partial}{\partial x_i}\left(K_{ij}\frac{\partial h^2}{\partial x_j}\right) = 2\overline{Q} \qquad (11\text{-}35)$$

References

1. Bear, J. & Verruijt, A., *Modeling groundwater flow and pollution*. Kluwer Academic Publ., Dordrecht, 1987.
2. Diersch, H.-J. G., *FEFLOW - Reference Manual*. WASY Ltd., Berlin, 2002.
3. Illman, W. A., Thompson, D. L., Vesselinov, V. V., Chen, G. & Neuman, S. P., *Single- and cross-hole pneumatic tests in unsaturated fractured tuffs at the Apache Leap Research Site: Phenomenology, spatial variability, connectivity and scale*. NUREG/CR-5559, 187p., U.S. Nuclear Regulatory Commission, 1998.
4. Shan, C., Falta, R.W. & Javandel, I., Analytical solutions for steady state gas flow to a soil vapor extraction well. *Water Resour. Res.* **28** (1992) 4, 1105-1120.

11. Remarks on gas flow modeling by using FEFLOW

Steady-state linearized Richards equation for fast solution of unsaturated flow systems (FUSY)

12

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

[a]*Centre d'Hydrogéologie, Université de Neuchâtel, Switzerland*
[b]*WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany*

12.1 Motivation

It is well-know the solution of the nonlinear Richards equation for unsaturated flow require an increased numerical effort and convergence difficulties can occur at dry conditions[3]. Considering complex field situations and large three-dimensional (3D) applications the computational burden can be extremely high. When there are only interests in steady-state solutions, e.g. capillary barrier problems[3], the standard approach in form any time-stepping strategy becomes very circumstantial and inefficient. For large and complicated problems the nonlinearity of the Richards equation represents serious limitations and solution alternatives are needed.

12.2 Basic Balance Equations

The mass conservation equation of a fluid in an unsaturated media is for steady-state conditions (symbols are listed in the Appendix A 'Nomenclature')

$$\nabla \cdot \boldsymbol{q} = Q \qquad (12\text{-}1)$$

The fluid motion (fluid momentum balance) is described by the Darcy equation written in the form

$$\boldsymbol{q} = -K_r(\psi)\boldsymbol{K}[\nabla\psi + (1+\chi)\boldsymbol{e}] \qquad (12\text{-}2)$$

Inserting (12-2) into (12-1) the *Richards equation* for a steady-state flow results

$$-\nabla \cdot \{\boldsymbol{K}K_r(\psi)[\nabla\psi + (1+\chi)\boldsymbol{e}]\} = Q \qquad (12\text{-}3)$$

which has to be solved for ψ. The steady Richards equation (12-3) is nonlinear due to the dependence of the relative hydraulic conductivity K_r on the pressure head ψ. For $K_r(\psi)$ an appropriate constitutive relationship is required.

12.3 Exponential Law of Relative Hydraulic Conductivity

There are different constitutive relationships of the relative hydraulic conductivity $K_r(\psi)$. Most common

are the van Genuchten and the Brooks-Corey parametric models[3]. An other alternative is the *exponential law* in the form

$$K_r(\psi) = \begin{cases} e^{\alpha\psi} & \text{for} \quad \psi < 0 \\ 1 & \text{for} \quad \psi \geq 0 \end{cases} \qquad (12\text{-}4)$$

It can be shown that the exponential law (12-4) for the relative hydraulic conductivity give results similar to the standard van Genuchten and the Brooks-Corey parametric models. It is to be emphasized that for steady-state conditions no assumption is required for the relationship between pressure head ψ and saturation s. In this context, the ψ-values can be transformed into saturations using any curve, like for example the van Genuchten relationship. Accordingly, *mixed parametric models* can be a promising alternative for solving complex unsaturated flow problems under steady-state conditions, where for example the exponential law is used for $K_r(\psi)$ and the van Genuchten law is used to find the saturations $s(\psi)$.

> ⚠ **Notice,** *the FUSY approach is restricted to homogeneous soils, where the α-parameter appearing in exponential law for the relative conductivity (12-4) is a constant.*

12.4 Transformation of the Steady-state Richards Equation

Introducing the following *Kirchhoff integral transform* related to the relative hydraulic conductivity in the form

$$F(\psi) = \int_{-\infty}^{\psi} K_r(\tau)d\tau \qquad (12\text{-}5)$$

one gets with the exponential law (12-4)

$$F(\psi) = \int_{-\infty}^{\psi} e^{\alpha\tau}d\tau = \frac{1}{\alpha}e^{\alpha\tau}\Big|_{-\infty}^{\psi} = \frac{1}{\alpha}e^{\alpha\psi} = \frac{1}{\alpha}K_r(\psi) \quad (12\text{-}6)$$

From (12-6) it follows

$$\nabla F = e^{\alpha\psi}\nabla\psi = K_r(\psi)\nabla\psi \qquad (12\text{-}7)$$

and

$$K_r(\psi) = \alpha F \qquad (12\text{-}8)$$

Important note: The expression (12-7) is only valid if assuming a constant α-parameter. This restricts the FUSY approach to problems with unsaturated parameters α being spatially invariable.

Taking the Richards equation (12-3)

$$-\nabla \cdot \left[\boldsymbol{K} \cdot \underbrace{K_r(\psi)\nabla\psi}_{\nabla F} + \boldsymbol{K} \cdot \underbrace{K_r(\psi)}_{\alpha F}(1+\chi)\boldsymbol{e} \right] = Q \quad (12\text{-}9)$$

the transformation leads to a *steady-state advection-diffusion equation* in the form

$$-\nabla \cdot (\boldsymbol{K} \cdot \nabla F - \boldsymbol{v}F) = Q \qquad (12\text{-}10)$$

for

$$F \in \left(0, \frac{1}{\alpha}\right) \qquad (12\text{-}11)$$

with the advective velocity vector

$$v = -\alpha K \cdot (1 + \chi)e \qquad (12\text{-}12)$$

Notice, the advection-diffusion equation (12-10) is *linear* in F and represents a *divergence-form type*[1].

The solution of such a type of linear advection-diffusion equation, in form of a *steady-state linearized Richards equation*, can be performed very efficiently and fast by using standard techniques. It is denoted by the acronym *FUSY (fast solution of unsaturated flow systems)*.

After solving (12-10) for F in the domain Ω and for given boundary conditions on $\partial\Omega$ the pressure head ψ can be simply obtained. Since

$$\alpha F = e^{\alpha\psi} \qquad (12\text{-}13)$$

and

$$\ln(\alpha F) = \alpha\psi \qquad (12\text{-}14)$$

one uses

$$\psi = \frac{1}{\alpha}\ln(\alpha F) \qquad (12\text{-}15)$$

for the retransformation. As soon as ψ is known the saturation s can be determined from any $s(\psi)$ curve. Additionally, the hydraulic head (potential) h is given by

$$h = \psi + z = \frac{1}{\alpha}\ln(\alpha F) + z \qquad (12\text{-}16)$$

and the resulting Darcy fluxes q are computed by

$$q = -\underbrace{K_r(\psi)}_{e^{\alpha\psi}} K(\nabla h + \chi e) \qquad (12\text{-}17)$$

12.5 Boundary Conditions

For the domain Ω boundary conditions (BC) are imposed on the four disjoint portions of the total boundary $\partial\Omega$:

$$\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4 \qquad (12\text{-}18a)$$

with

$$\Gamma_2 = \Gamma_2^a \cup \Gamma_2^b \qquad (12\text{-}18b)$$

12.5.1 Untransformed conditions

For the untransformed Richards-type model equation (12-3) boundary conditions are prescribed by the following types[2]:

Dirichlet 1st kind BC for hydraulic head

$$h = h_1 \qquad \text{on} \qquad \Gamma_1 \qquad (12\text{-}19)$$

Neumann flux-type 2nd kind BC

$$q_n = -[K_r(\psi)K \cdot (\nabla h + \chi e)] \cdot n \qquad \text{on} \qquad \Gamma_2^a \qquad (12\text{-}20)$$

Neumann gradient-type 2nd kind BC
(a special type for a free drainage at diminishing pressure gradient $\nabla\psi \approx 0$)

$$q_n = -[\boldsymbol{K} \cdot (1 + \chi)\boldsymbol{e}] \cdot \boldsymbol{n} \qquad \text{on} \qquad \Gamma_2^b \qquad (12\text{-}21)$$

Cauchy 3rd kind BC

$$q_n = -\Phi(h_2 - h) \qquad \text{on} \qquad \Gamma_3 \qquad (12\text{-}22)$$

Seepage face BC's
(represents a flux-constrained Dirichlet boundary-value problem)

$$\begin{aligned} h &= z \\ \psi &= 0 \end{aligned} \qquad \text{at} \qquad q_n > 0 \qquad\qquad (12\text{-}23)$$

$$\text{(flux is going outward)} \qquad \text{on} \qquad \Gamma_4$$

12.5.2 Transformed conditions

Boundary conditions for the governing advection-diffusion equation (12-10) can be equivalently found for (12-19), (12-20), (12-21), (12-22) and (12-23) when written by the new F variable:

Dirichlet 1st kind BC for hydraulic head

$$F = \frac{1}{\alpha}e^{\alpha(h_1 - z)} \qquad \text{on} \qquad \Gamma_1 \qquad (12\text{-}24)$$

Neumann flux-type 2nd kind BC

$$q_n = -\left[\boldsymbol{K} \cdot \underbrace{K_r(\psi)\nabla\psi}_{\nabla F} + \boldsymbol{K} \cdot \underbrace{K_r(\psi)}_{\alpha F}(1+\chi)\boldsymbol{e}\right] \cdot \boldsymbol{n} \quad (12\text{-}25)$$

$$= -(\boldsymbol{K} \cdot \nabla F - \boldsymbol{v}F) \cdot \boldsymbol{n} \qquad \text{on} \qquad \Gamma_2^a$$

Neumann gradient-type 2nd kind BC
(free drainage at diminishing pressure gradient, i.e., $\nabla\psi = \nabla F \approx 0$)

$$q_n = -\underbrace{[\boldsymbol{K} \cdot (1+\chi)\boldsymbol{e}]}_{-\boldsymbol{v}/\alpha} \cdot \boldsymbol{n} = \frac{\boldsymbol{v}}{\alpha} \cdot \boldsymbol{n} \quad \text{on} \quad \Gamma_2^b \quad (12\text{-}26)$$

Cauchy 3rd kind BC

$$q_n = -\Phi(h_2 - h) = -\Phi\left[h_2 - z - \frac{1}{\alpha}\ln(\alpha F)\right] \qquad (12\text{-}27)$$

$$\text{on} \quad \Gamma_3$$

which represents a nonlinear expression in F .

Seepage face BC's

$$F = \frac{1}{\alpha} \qquad \text{at} \qquad q_n > 0 \qquad\qquad (12\text{-}28)$$

$$\text{(flux is going outward)} \qquad \text{on} \qquad \Gamma_4$$

References

1. Diersch, H.-J.G., About the difference between the convective and the divergence form of the transport equation, *White Papers - Vol. I,* Chapter 6, WASY Ltd., 2002.

2. Diersch, H.-J.G., Treatment of free surfaces in 2D and 3D groundwater modelling, *Math. Geologie* **2** (1998), 17-43.

3. Diersch, H.-J.G. & Perrochet, P., On the primary variable switching technique for simulating unsaturated-saturated flows, *Advances in Water Resources* **23** (1999) 3, 271-301.

Appendix A

Nomenclature

In the above the symbols have the following meaning:

<u>Latin symbols</u>

e	=	gravitational unit vector, (1);
F	=	transform function, (L);
h	=	$\psi + z$, hydraulic (piezometric) head, (L);
h_1, h_2	=	prescribed boundary values for h, (L);
$K_r(s)$	=	relative hydraulic conductivity ($0 < K_r \leq 1$, $K_r = 1$ if saturated at $s = 1$), (1);
K	=	tensor of hydraulic conductivity for the saturated medium (anisotropy), (LT^{-1});
n	=	normal unit vector (positive outward);
Q	=	fluid mass sink/source, (T^{-1});
q	=	Darcy flux vector, (LT^{-1});
q_n	=	normal flux on a boundary (directed positive outward), (LT^{-1});
$s(\psi)$	=	saturation, $(0 < s \leq 1$, $s = 1$ if medium is saturated), (1);

z	=	elevation above a reference datum, (L);
v	=	advective velocity vector, (T^{-1});

<u>Greek symbols</u>

α	=	sorptive number, (L^{-1});
Γ	=	boundary;
$\partial\Omega$	=	total boundary;
ε	=	porosity (= volume fraction of fluid phase), (1);
τ	=	integration variable, (L);
Φ	=	transfer coefficient, (T^{-1});
χ	=	buoyancy coefficient including fluid density effects, (1);
ψ	=	pressure head (L);
Ω	=	domain;
∇	=	Nabla (vector) operator, (L^{-1});

<u>Subscripts</u>

e	elemental;
i, j, k	nodal or spatial indices;
o	reference value;

<u>Superscripts</u>

f	fluid (water) phase;
s	solid phase;

12. Steady-state linearized Richards equation for fast solution of unsaturated flow systems (FUSY)

The Petrov-Galerkin least square method (PGLS) 13

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

13.1 Introduction

Both in solving the nonlinear Richards equation and the advective-dispersive transport equations for mass and heat the advective part often becomes dominant and classic numerical techniques can completely fail. Conventionally, for multi-dimensional transport problems upwind techniques such as streamline-upwind Petrov-Galerkin (SUPG) or scalar upstream weighting are standard to stabilize the solutions. While the classic artificial diffusion method often suffers in a considerable smearing of steep fronts the SUPG formulation cannot preclude the presence of overshoots and undershoots in the vicinity of sharp gradients[10]. More recently, nonlinear shock-capturing techniques have been developed and successfully applied[5]. Otherwise, for unsaturated problems upstream weighting techniques can be helpful to damp out wiggles in the saturation profiles[6]. Unfortunately, most of these established techniques reveal over-diffusive properties and there is a further demand for alternative higher-order upwind techniques with reduced spurious numerical dispersion.

The Petrov-Galerkin least square (PGLS) finite-element method appears as a promising technique for tackling advective-dominant flow and transport processes at variably saturated conditions. Basic work of the PGLS was presented by Nguyen and Reynen[15]. In the context of the finite-element method (FEM) is based on a Petrov-Galerkin weak formulation where a 'modified' weighting function is derived from the least-squares finite element concept. This procedure is equivalent to the methods developed by Hughes and Brooks[10] and by Kelly *et al.*[12] where an artificial diffusion is added to the physical parameter. However, in this method the artificial diffusion can be derived directly from the least-squares finite element concept and requires no 'free' parameter. PGLS is recognized as a member of more general *stabilized finite element methods* as described by Codina[3].

13.2 The PGLS Approach Based on a Symmetric Streamline Stabilization (S3) via an Operator Splitting Technique

13.2.1 Operator split

Rather independently, a more general solution approach which is comparable to the PGLS technique as proposed by Nguyen and Reynen[15] has been developed by König[13] and Wendland[18]. König[13] used the operator split (OS) technique to solve the transport equations in a two-pass strategy, where the separate equations for the diffusive and the advective part are solved successively. Wendland[18] enhanced the OS technique by introducing a powerful one-pass approach termed as *symmetric streamline stabilization* (S3), where the diffusive and advective parts are re-assembled in one symmetric matrix system. A detailed discussion of the S3 approach can be found in Wendland[18]. In contrast to Nguyen and Reynen's PGLS strategy[15] the S3 strategy provides a more generalized approach with respect to temporal discretization techniques, where different time stepping schemes, including adaptive predictor-corrector strategies typically used in FEFLOW[4] can be easily covered.

Let us start with the following mass transport equation

$$(\varepsilon + \kappa)\dot{C} + \varepsilon \boldsymbol{v} \cdot \nabla C - \nabla \cdot (\boldsymbol{D} \cdot \nabla C) + \lambda C = f \quad (13\text{-}1)$$

where ε is the porosity, κ is the adsorption, $\boldsymbol{v}$ is the pore velocity vector, $\boldsymbol{D}$ is the 'diffusion' tensor, λ is the reaction and f is a sink/source term,

which can be written in the operator-splitted form

$$\underbrace{(\varepsilon + \kappa)}_{R_d}\dot{C} + (L_1 + L_2)C = f \quad (13\text{-}2)$$

with

$$L_1 = -\nabla \cdot (\boldsymbol{D} \cdot \nabla) + \lambda \quad (13\text{-}3a)$$

$$L_2 = \varepsilon \boldsymbol{v} \cdot \nabla \quad (13\text{-}3b)$$

Consider the solution $C(t)$ within the finite time interval $(t_n, t_n + \Delta t_n)$, where Δt_n denotes the variable timestep length and n indicates the time plane. The unknown function $C(t)$ is defined as

$$C^n = C(t_n) \quad (13\text{-}4)$$

at the old time plane, and

$$C^{n+1} = C(t_n + \Delta t_n) \quad (13\text{-}5)$$

at the new time plane. Furthermore, we introduce a time weighting factor $(0 \le \theta \le 1)$ in such a form

$$\left. \begin{array}{l} C(t_n + \theta \Delta t_n) = \theta C(t_n + \Delta t_n) + (1 - \theta)C(t_n) \\ C^{n+\theta} = \theta C^{n+1} + (1 - \theta)C^n \end{array} \right\} \quad (13\text{-}6)$$

In specifying θ we obtain different time discretization schemes, *viz.*,

$$\theta = 0 \quad \text{explicit scheme}$$
$$\theta = 1/2 \quad \text{trapezoid rule}$$
$$\theta = 1 \quad \text{implicit scheme}$$

(13-7)

Practically, from reasons of numerical stability we are only interested in semi-implicit and fully implicit schemes that means $1/2 \leq \theta \leq 1$.

The time derivatives are given as

$$\dot{C}^{n+\theta} = \dot{C}^{n+1} = \frac{C^{n+1} - C^n}{\Delta t_n}$$

(13-8)

for the backward Euler (BE) fully implicit scheme and

$$\dot{C}^{n+1} = \frac{2}{\Delta t_n}(C^{n+1} - C^n) - \dot{C}^n$$

(13-9)

for the trapezoid rule (TR) scheme. The TR expression can simplified if using a forward difference approximation for $\dot{C}^n = (C^{n+1} - C^n)/\Delta t_n$ which leads to the same expression of Eq. (13-8) for $\dot{C}^{n+1}$ which is used for the 2nd order accurate Crank-Nicolson approximation.

Now, we split the solution C into the diffusive and the advective part such that

$$C = C_1 + C_2$$

(13-10)

and

$$\varepsilon \dot{C}_1 + \kappa \dot{C} + L_1 C = f$$

(13-11)

represents the diffusive system while

$$\varepsilon(\dot{C} - \dot{C}_1) + L_2 C = 0$$

(13-12)

is the purely advective system. Summing (13-11) and (13-12) we realize the original balance equation (13-2). The idea of the operator splitting technique is in approximating the diffusive and advective equations (13-11) and (13-12), respectively, in a separate manner. After that the total discrete balance equation is obtained by the assembly of these two parts. In doing this, we assume that the initial conditions for the diffusive and advective variables are

$$t = t^n \qquad C_1^n = C^n \qquad C_2^n = 0 \qquad \dot{C}_2^n = 0 \quad (13\text{-}13)$$

It is to be noted that the diffusive solution C_1 can be considered as an intermediate solution which represents a temporally discrete interpolation between the old and the new time level as evidenced in Fig. 13.1.
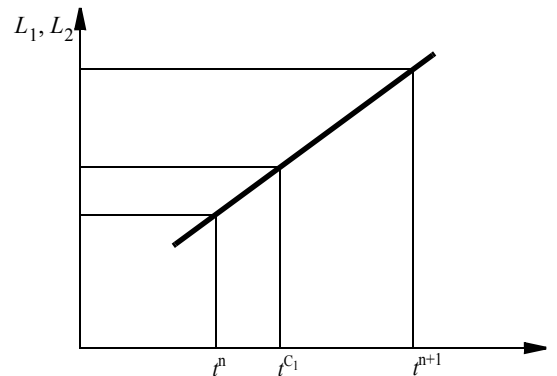
Figure 13.1 Temporally discrete interpolation of the intermediate solution C_1.

13.2.2 Approximation of the diffusive part

In the context of the FEM the unknown variables C, C_1 are replaced by a continuous approximation that assumes the separability of space and time, thus

$$
\left.\begin{aligned}
C(x_i, t) &\approx \sum_I N_I(x_i) C_I(t) \\
C_1(x_i, t) &\approx \sum_I N_I(x_i) C_{1I}(t)
\end{aligned}\right\} \tag{13-14}
$$

where the subscript $I = 1, ..., M$ denotes the nodal indices and $i = 1, ..., D, \; D \in (2, 3)$ represents the spatial indices.

For the diffusive part the standard Galerkin approximation is used. Thus, the residual $\Re$ of Eq. (13-11) as

$$
\Re = \varepsilon \dot{C}_1 + \kappa \dot{C} + L_1 C - f \tag{13-15}
$$

yields the following weak formulation

$$
\left.\begin{aligned}
\int_\Omega N_I \Re \, d\Omega &= 0 \\
\int_\Omega N_I(\varepsilon \dot{C}_1 + \kappa \dot{C} + L_1 C - f) d\Omega &= 0
\end{aligned}\right\} \tag{13-16}
$$

Using the trial space expression (13-14) the following global matrix system results

$$
M \cdot \dot{C}_1 + S \cdot \dot{C} + B \cdot C = P \tag{13-17}
$$

with its components written in indicial notation

$$
M_{IJ} = \int_\Omega \varepsilon N_I N_J d\Omega \tag{13-18a}
$$

$$
S_{IJ} = \int_\Omega \kappa N_I N_J d\Omega \tag{13-18b}
$$

$$
B_{IJ} = \int_\Omega \frac{\partial N_I}{\partial x_i} D_{ij} \frac{\partial N_J}{\partial x_j} d\Omega + \int_\Omega \lambda N_I N_J d\Omega \tag{13-18c}
$$

$$
P_I = \int_\Omega N_I f d\Omega - \int_\Gamma N_I q_c d\Gamma \tag{13-18d}
$$

where the subscripts $I, J = 1, ..., M$ denote the nodal indices and $i, j = 1, ..., D$ represent the spatial indices. Furthermore, q_c corresponds to the outer diffusive flux on the boundary Γ.

The time discretization of Eq. (13-17) leads to the following expression

$$
\begin{aligned}
M \cdot \dot{C}_1^{n+\theta} + S \cdot \dot{C}_1^{n+\theta} + S \cdot \dot{C}_2^n + B \cdot C_1^{n+\theta} \\
+ B \cdot C_2^n = P^{n+\theta}
\end{aligned} \tag{13-19}
$$

With (13-13) we get the solution of the intermediate (diffusive) part C_1 at the new time plane $n+1$ as

$$
\begin{aligned}
(M + S) \cdot \left(\frac{C_1^{n+1} - C^n}{\Delta t_n} \right) + \theta B \cdot C_1^{n+1} \\
+ (1 - \theta) B \cdot C^n = P^{n+\theta}
\end{aligned} \tag{13-20}
$$

or

$$\left[\frac{(M+S)}{\Delta t_n} + \theta B\right] C_1^{n+1}$$
$$= \left[\frac{(M+S)}{\Delta t_n} - (1-\theta)B\right] C^n + P^{n+\theta}$$

(13-21)

For the predictor-corrector strategy based on the BE and TR schemes we obtain alternatively

$$\left[\frac{\sigma(M+S)}{\Delta t_n} + B\right] C_1^{n+1}$$
$$= (M+S)\cdot\left[\frac{\sigma}{\Delta t_n}C^n + (\sigma-1)\dot{C}^n\right] + P^{n+1}$$

(13-22)

with

$$\sigma = \frac{1}{\theta} \qquad \theta \geq \frac{1}{2}$$

(13-23)

13.2.3 Approximation of the advective part

The residual of the advective part (13-12) in form of

$$\left.\begin{array}{l}\Im = \varepsilon(\dot{C} - \dot{C}_1) + L_2 C \\ \Im = \varepsilon(\dot{C} - \dot{C}_1) + \varepsilon v \cdot \nabla C\end{array}\right\}$$

(13-24)

will be handled by the least-square method

$$\frac{\partial}{\partial C_I}\int_\Omega (\Im \cdot \Im)d\Omega = 0$$

(13-25)

This is equivalent to the minimization process in using the weighting (test) function in form of

$$\psi_I = \frac{\partial\Im}{\partial C_I} = \frac{\varepsilon N_I}{\Delta t_n} + \theta\varepsilon v \cdot \nabla N_I$$

(13-26)

for the θ-based time marching schemes and

$$\psi_I = \frac{\partial\Im}{\partial C_I} = \frac{\sigma\varepsilon N_I}{\Delta t_n} + \varepsilon v \cdot \nabla N_I$$

(13-27)

for the BE and TR predictor-corrector approximations.

The weak formulation yields

$$\int_\Omega \psi_I \Im d\Omega = 0$$

(13-28)

and is exemplified for the θ-based time marching scheme as

$$\int_\Omega (N_I + \theta\Delta t_n v \cdot \nabla N_I)[\varepsilon(\dot{C} - \dot{C}_1) + \varepsilon v \cdot \nabla C]d\Omega = 0 \quad (13\text{-}29)$$

where the residual is weighted by the test function consisting of two parts as displayed in Fig. 13.2.

$$N_I \qquad\qquad \theta\Delta t_n v \cdot \nabla N_I \qquad\qquad \text{LS upwind weighting}$$

Figure 13.2 Least-square test function of the operator split (modified from [13]).

13. The Petrov-Galerkin least square method (PGLS)

The least-square weak statement (13-29) leads to the following matrix system

$$(M + \theta \Delta t_n V) \cdot \dot{C} + (K + \theta T) \cdot C \qquad (13\text{-}30)$$
$$= (M + \theta \Delta t_n V) \cdot \dot{C}_1$$

with the components

$$V_{IJ} = \int_\Omega \varepsilon(v \cdot \nabla N_I) N_J d\Omega \qquad (13\text{-}31a)$$

$$K_{IJ} = \int_\Omega \varepsilon N_I(v \cdot \nabla N_J) d\Omega \qquad (13\text{-}31b)$$

$$T_{IJ} = \int_\Omega \Delta t_n \varepsilon(v \cdot \nabla N_I) \cdot (v \cdot \nabla N_J) d\Omega \qquad (13\text{-}31c)$$

Analogously, for the BE and TR predictor-corrector technique we have to apply

$$\int_\Omega (\sigma N_I + \Delta t_n v \cdot \nabla N_I)[\varepsilon(\dot{C} - \dot{C}_1) + \varepsilon v \cdot \nabla C]d\Omega = 0 \ (13\text{-}32)$$

or with (13-23)

$$\int_\Omega (N_I + \theta \Delta t_n v \cdot \nabla N_I)[\varepsilon(\dot{C} - \dot{C}_1) + \varepsilon v \cdot \nabla C]d\Omega = 0 \ (13\text{-}33)$$

which is the same as for θ-based time marching scheme which gives the matrix system (13-30).

The time discretization of Eq. (13-30) leads to the following expression

$$(M + \theta \Delta t_n V) \cdot \dot{C}^{n+\theta} + (K + \theta T) \cdot C^{n+\theta} \qquad (13\text{-}34)$$
$$= (M + \theta \Delta t_n V) \cdot \dot{C}_1^{n+\theta}$$

We get the solution of the advective part at the new time plane $n+1$ as

$$(M + \theta \Delta t_n V) \cdot \left(\frac{C^{n+1} - C^n}{\Delta t_n} \right) + \theta(K + \theta T) \cdot C^{n+1} + (1 - \theta)(K + \theta T) \cdot C^n =$$
$$(13\text{-}35)$$
$$= (M + \theta \Delta t_n V) \cdot \left(\frac{C_1^{n+1} - C^n}{\Delta t_n} \right)$$

and finally

$$\left[\frac{M}{\Delta t_n} + \theta(V + K + \theta T) \right] C^{n+1} \qquad (13\text{-}36)$$
$$= \left(\frac{M}{\Delta t_n} + \theta V \right) C_1^{n+1} - (1 - \theta)(K + \theta T) \cdot C^n$$

Regarding the predictor-corrector strategy based on the BE and TR schemes if taking

$$\left. \begin{aligned} \dot{C}^{n+1} &= \frac{\sigma}{\Delta t_n}(C^{n+1} - C^n) - (\sigma - 1)\dot{C}^n \\[2mm] \dot{C}_1^{n+1} &= \frac{\sigma}{\Delta t_n}(C_1^{n+1} - C^n) - (\sigma - 1)\dot{C}^n \end{aligned} \right\} \qquad (13\text{-}37)$$

and using (13-30) we obtain alternatively for the advective part the following matrix system

$$\left(\frac{\sigma M}{\Delta t_n} + V + K + \frac{1}{\sigma}T \right) C^{n+1} = \left(\frac{\sigma M}{\Delta t_n} + V \right) C_1^{n+1} \qquad (13\text{-}38)$$

13.2.4 Assembly of the diffusive and advective parts

Finally, to obtain the matrix system for the total balance equation (13-1) we have to add the diffusive and advective parts.

For the θ-based time marching scheme the summation of (13-36) and (13-21) yields

$$\left[\frac{M}{\Delta t_n} + \theta(V + K + \theta T)\right]C^{n+1} + \left[\frac{(M+S)}{\Delta t_n} + \theta B\right]C_1^{n+1} =$$
$$= \left(\frac{M}{\Delta t_n} + \theta V\right)C_1^{n+1} - (1-\theta)(K+\theta T)\cdot C^n + \left[\frac{(M+S)}{\Delta t_n} - (1-\theta)B\right]C^n + P^{n+\theta}$$

(13-39)

The term $\frac{M}{\Delta t_n}C_1^{n+1}$ can be eliminated from (13-39). The remaining terms correlated with the intermediate solution C_1^{n+1} can be transformed in the following way[18]. All terms with C_1^{n+1} on the left-hand side are replaced by C^{n+1}, while such terms on the right-hand side are substituted by C^n. In doing so, the following matrix system results

$$\left[\frac{(M+S)}{\Delta t_n} + \theta(B + V + K + \theta T)\right]C^{n+1} =$$
$$= \left[\frac{(M+S)}{\Delta t_n} - (1-\theta)(B + K + \theta T) + \theta V\right]C^n + P^{n+\theta}$$

(13-40)

or

$$\left[\frac{O}{\Delta t_n} + \theta(B + V + K + \theta T)\right]C^{n+1}$$
$$= \left[\frac{O}{\Delta t_n} - (1-\theta)(B + K + \theta T) + \theta V\right]C^n + P^{n+\theta}$$

(13-41)

with

$$O_{IJ} = \int_\Omega R_d N_I N_J d\Omega$$

$$B_{IJ} = \int_\Omega \frac{\partial N_I}{\partial x_i} D_{ij} \frac{\partial N_J}{\partial x_j} d\Omega + \int_\Omega \lambda N_I N_J d\Omega$$

$$V_{IJ} = \int_\Omega (q \cdot \nabla N_I) N_J d\Omega$$

$$K_{IJ} = \int_\Omega N_I (q \cdot \nabla N_J) d\Omega$$

$$T_{IJ} = \int_\Omega \frac{\Delta t_n}{\varepsilon}(q \cdot \nabla N_I) \cdot (q \cdot \nabla N_J) d\Omega$$

$$P_I = \int_\Omega N_I f d\Omega - \int_\Gamma N_I q_c d\Gamma$$

(13-42)

where $R_d = \varepsilon + \kappa$ is the retardation factor and $q = \varepsilon v$ is the Darcy velocity vector.

Analogously for the BE and TR predictor-corrector schemes we add (13-38) and (13-22)

$$\left(\frac{\sigma M}{\Delta t_n} + V + K + \frac{1}{\sigma}T\right)C^{n+1} + \left[\frac{\sigma(M+S)}{\Delta t_n} + B\right]C_1^{n+1} =$$
$$= \left(\frac{\sigma M}{\Delta t_n} + V\right)C_1^{n+1} + (M+S)\cdot\left[\frac{\sigma}{\Delta t_n}C^n + (\sigma-1)\dot{C}^n\right] + P^{n+1}$$

(13-43)

which gives

$$\left[\frac{\sigma O}{\Delta t_n} + B + V + K + \frac{1}{\sigma}T\right]C^{n+1}$$
$$= VC^n + O\cdot\left[\frac{\sigma}{\Delta t_n}C^n + (\sigma-1)\dot{C}^n\right] + P^{n+\theta}$$

(13-44)

The final matrix systems (13-41) and (13-44) for the θ-based time marching scheme and the BE and TR predictor-corrector scheme, respectively, are symmetric and positive definite. This results from the fact that the advective matrices V and K form a symmetric contribution as the sum $(V + K)$ because $K = V^T$ is the transpose as easily seen from (13-42). The symmetric term T can be interpreted as an additional term of artificial diffusion. This naturally arises from the least-square weighting procedure (13-25).

13.2.5 Remarks on the application of operator split to the divergence form of the transport equation

FEFLOW[4] differs between the convective and the divergence forms in solving the governing transport equations. The advantage of the divergence form is in prescribing a total (advective plus diffusive) boundary flux on Γ instead of imposing only the diffusive flux q_c (*cf.* Eq. (13-18d) on the boundary Γ. The balance equation (13-1) represents a convective form which is standard, where the continuity equation is explicitly used to find a gradient-type relationship $\varepsilon v \cdot \nabla C$ for the advective term. But basically, the balance equation possesses a divergence expression $\nabla \cdot (\varepsilon v C)$. It reads

$$(\varepsilon + \kappa)\dot{C} + \nabla \cdot (\varepsilon v C) - \nabla \cdot (D \cdot \nabla C) + \lambda C = f \quad (13\text{-}45)$$

Then, the splitted advective part yields

$$\varepsilon(\dot{C} - \dot{C}_1) + L_2 C = 0 \quad (13\text{-}46)$$

with the advective operator

$$L_2 = \varepsilon v \cdot \nabla + (\nabla \cdot \varepsilon v) \quad (13\text{-}47)$$

The weak statement of the least-square weighting method is then

$$\int_\Omega \left[N_I + \theta \Delta t_n \frac{1}{\varepsilon}\nabla \cdot (\varepsilon v N_I)\right][\varepsilon(\dot{C} - \dot{C}_1) + \nabla \cdot (\varepsilon v C)]d\Omega = 0 \quad (13\text{-}48)$$

which leads to the following matrix system

$$(M + \theta \Delta t_n V^*) \cdot \dot{C} + (K^* + \theta T^*) \cdot C \quad (13\text{-}49)$$
$$= (M + \theta \Delta t_n V^*) \cdot \dot{C}_1$$

with the components

$$V_{IJ}^* = \int_\Omega \varepsilon(v \cdot \nabla N_I)N_J d\Omega + \int_\Omega N_I N_J \nabla \cdot (\varepsilon v)d\Omega \quad (13\text{-}50a)$$

$$K_{IJ}^* = -\int_\Omega \varepsilon v N_J \cdot \nabla N_I d\Omega + \int_\Gamma N_I q_c^a d\Gamma \quad (13\text{-}50b)$$

$$T_{IJ}^* = \int_\Omega \Delta t_n \frac{1}{\varepsilon}[\nabla \cdot (\varepsilon v N_I)] \cdot [\nabla \cdot (\varepsilon v N_J)]d\Omega \quad (13\text{-}50c)$$

where we transformed the partial advective matrix K_{IJ}^* by using the divergence theorem according to

$$\int_\Omega N_I \nabla \cdot (\varepsilon v C)d\Omega = \int_\Omega \nabla \cdot (N_I \varepsilon v C)d\Omega \quad (13\text{-}51)$$
$$- \int_\Omega \varepsilon v C \cdot \nabla N_I d\Omega = \int_\Gamma N_I q_c^a d\Gamma - \int_\Omega \varepsilon v C \cdot \nabla N_I d\Omega$$

to get the required description of the advective outer border flux q_c^a.

The time discretization of (13-49) for the divergence form of the splitted advective part yields finally

$$\left[\frac{M}{\Delta t_n} + \theta(V^* + K^* + \theta T^*)\right]C^{n+1}$$ (13-52)
$$= \left(\frac{M}{\Delta t_n} + \theta V^*\right)C_1^{n+1} - (1-\theta)(K^* + \theta T^*) \cdot C^n$$

Inspecting Eq. (13-52) with its matrix components (13-50a) to (13-50c) it reveals the following. The symmetric property of the matrix system is lost because K^* is no more a transpose of V^*. Furthermore, the evaluation of the divergence expressions $\nabla \cdot (\varepsilon v)$ in (13-50a) and $\nabla \cdot (\varepsilon v N_I)$ in (13-50c) causes difficulties if the flow is not *solenoidal* (divergence-free) at the presence of compression and sink/sources of fluid. As the sum, it becomes apparent that the least-square technique is rather inappropriate for the divergence form of the transport equation.

13.3 Integration of PGLS into the FEFLOW Simulation Package

13.3.1 General

In FEFLOW various numerical schemes are available for flow and transport processes in variably saturated media. Regarding the 2D and 3D transport equations for solute and heat it covers:

(1) GFEM-Bubnov-Galerkin FEM (GFEM)
(2) SU-Petrov-Galerkin streamline upwind

(3) FU-Full upwind technique
(4) SC-Shock-capturing technique.

With respect to the unsaturated flow equations in two and three dimensions FEFLOW is currently capable of performing

- Standard h-based Richards equation form via either Picard or Newton iteration;
- Mixed $\psi - s$-based Richards equation by using either a modified Picard scheme or the Newton method;
- Primary variable switching technique with a full Newton method;
- Upstream weighting based on a Gauss-point-related evaluation technique for the relative conductivity.

The PGLS scheme for both transport and unsaturated flow processes has been developed and implemented as the fifth option of numerical schemes in FEFLOW, *viz.*,

(5) PGLS-Petrov-Galerkin least square method.

The main advantages of the PGLS can be summarized as follows:

- It is compatible to the existing techniques within the finite element context.
- It represents a higher-order upwind scheme providing an improved temporal and spatial accuracy.
- The scheme provides a built-in upwind characteristics; accordingly, no additional 'free' parameters have to be specified by the user.
- It leads to symmetric matrix systems, which can be solved effectively and fast. Compared to the

conventional unsymmetric approaches the PGLS is a cost-effective variant with a significantly reduced storage demand.

- It can be simply applied both to the transport equations and to the unsaturated flow equations.

13.3.2 Resumé of basic equations

The following nonlinear system is solved in two and three dimensions (symbols are listed in Appendix A):

$$S_o \cdot s(\psi)\frac{\partial \psi}{\partial t} + \varepsilon\frac{\partial s(\psi)}{\partial t} + \nabla \cdot \boldsymbol{q} = Q_h \qquad (13\text{-}53)$$

$$\boldsymbol{q} = -K_r(s)\boldsymbol{K}f_\mu\left[\nabla\psi + \left(1 + \frac{\rho - \rho_o}{\rho_o}\right)\boldsymbol{e}\right] \qquad (13\text{-}54)$$

$$s(\psi)\varepsilon R_d(C)\frac{\partial C}{\partial t} + \boldsymbol{q}\cdot\nabla C$$
$$- \nabla\cdot[(\varepsilon s(\psi)D_d\boldsymbol{I} + \boldsymbol{D})\cdot\nabla C] \qquad (13\text{-}55)$$
$$+ [s(\psi)\varepsilon R(C)\vartheta + Q_h]C = s(\psi)Q_C$$

$$[s(\psi)\varepsilon\rho c^f + (1-\varepsilon)\rho^s c^s]\frac{\partial T}{\partial t} + \rho c^f\boldsymbol{q}\cdot\nabla T$$
$$- \nabla\cdot[(\Delta + (1-\varepsilon)\lambda^s\boldsymbol{I})\cdot\nabla T] \qquad (13\text{-}56)$$
$$+ \rho c^f Q_h(T-T_o) = Q_T(\psi)$$

with the definitions and constitutive relationships

$$h = \frac{p}{\rho_o g} + z = \psi + z \qquad (13\text{-}57a)$$

$$S_o = \varepsilon\gamma + (1-\varepsilon)\phi \qquad (13\text{-}57b)$$

$$\boldsymbol{K} = \frac{\boldsymbol{k}\rho_o g}{\mu_0} \qquad (13\text{-}57c)$$

$$f_\mu = \frac{\mu_0}{\mu(C,T)} \qquad (13\text{-}57d)$$

$$\rho = \rho_o[1 + \alpha(C-C_o) - \beta(T-T_o)] \qquad (13\text{-}57e)$$

$$\left.\begin{array}{l}R(C) = 1 + \dfrac{(1-\varepsilon)}{\varepsilon}\chi(C)\\[2mm] R_d(C) = 1 + \dfrac{(1-\varepsilon)}{\varepsilon}\dfrac{\partial[\chi(C)\cdot C]}{\partial C}\end{array}\right\} \qquad (13\text{-}57f)$$

$$\boldsymbol{D} = (\beta_L - \beta_T)\frac{\boldsymbol{q}\otimes\boldsymbol{q}}{\|\boldsymbol{q}\|} + \beta_T\|\boldsymbol{q}\|\boldsymbol{I} \qquad (13\text{-}57g)$$

$$\Delta = \rho c^f\boldsymbol{D} + \varepsilon s(\psi)\lambda^f\boldsymbol{I} \qquad (13\text{-}57h)$$

$$Q_T(\psi) = s(\psi)\varepsilon Q_T^f + (1-\varepsilon)Q_T^s \qquad (13\text{-}57i)$$

To solve the nonlinear equations (13-53) to (13-56) for ψ, s, $\boldsymbol{q}$, C, and T under unsaturated-saturated conditions constitutive relationships are additionally required for the saturation s as a function of the pressure (capillary) head ψ, with its inverse, the pressure head ψ as a function of the saturation s, and for the relative hydraulic conductivity K_r as a function of the pressure head ψ or the saturation s. FEFLOW provides the van Genuchten-Mualem, the Brooks-Corey, the Haverkamp, the exponential and the linear parametric models.

The above equations (13-53) to (13-56) are discretized by the FEM using bilinear or biquadratic elements for 2D, and prismatic pentahedral trilinear or hexahedral trilinear and triquadratic elements for 3D

(Fig. 13.3).

2D   3D
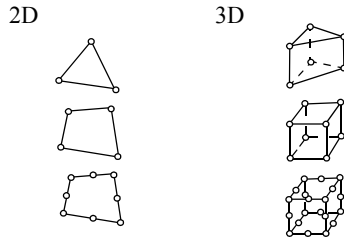
Figure 13.3 FEFLOW's element types.

Finally, it yields the following coupled matrix system:

$$
\left.
\begin{aligned}
O\dot{h} + S(h, C, T)h &= F(h, q, C, \dot{C}, T, \dot{T}) \\
Aq &= B(h, C, T) \\
P(C)\dot{C} + D(q, C)C &= R(C) \\
U\dot{T} + L(q, T)T &= W(T)
\end{aligned}
\right\}
\quad (13\text{-}58)
$$

where $h = h(\psi, s)$, q, C and T represent the resulting vectors of nodal hydraulic head (alternatively, pressure head or saturation), Darcy fluxes, contaminant concentration and temperature, respectively. The superposed dot means differentiation with respect to time t. The matrices S, A, O, P and U are symmetric and sparse, while the sparse D and L matrices are only symmetric for the PGLS scheme. The remaining vectors F, B, R and W encompass the right-hand sides (RHS) of eqns (13-53) to (13-56), respectively. The main functional dependence is shown in parenthesis.

REMARK 1. The matrices D and L are unsymmetric for

the standard schemes such as Bubnov-Galerkin (GFEM), streamline upwinding (SUPG) and shock capturing (SC). The matrix S also becomes unsymmetric for the primary variable switching technique[6] applied to unsaturated flow problems based on the Newton iteration technique.

The most remarkable features of the PGLS scheme applied to the system (13-58) are summarized for the mass transport equation as follows:
In using the different time discretization schemes in form of both (a) the θ-weighted time marching with fixed step sizes and (b) the predictor-corrector time stepping of first and second order in time the matrix equation

$$
P(C)\dot{C} + D(q, C)C = R(C) \quad (13\text{-}59)
$$

leads to the temporally discretized matrix relationship for the θ-weighted scheme as

$$
\left[\frac{P}{\Delta t_n} + \theta(B + V + K + \theta G)\right]C^{n+1} \quad (13\text{-}60)
$$
$$
= \left[\frac{P}{\Delta t_n} - (1-\theta)(B + K + \theta G) + \theta V\right]C^n + R^{n+\theta}
$$

and for the predictor-corrector scheme as

$$
\left[\frac{\sigma P}{\Delta t_n} + B + V + K + \frac{1}{\sigma}G\right]C^{n+1} \quad (13\text{-}61)
$$
$$
= VC^n + P \cdot \left[\frac{\sigma}{\Delta t_n}C^n + (\sigma - 1)\dot{C}^n\right] + R^{n+1}
$$

with

$$\sigma = \frac{1}{\theta} \qquad \theta \geq \frac{1}{2} \qquad (13\text{-}62)$$

The individual matrices for the PGLS scheme of (13-60) and (13-61) are given by

$$
\left.
\begin{aligned}
P_{IJ} &= \int_{\Omega} s(\psi) R_d N_I N_J d\Omega \\[4pt]
B_{IJ} &= \int_{\Omega} \nabla N_I \cdot \{[\varepsilon s(\psi) D_d \delta_{ij} + \boldsymbol{D}] \cdot \nabla N_J\} d\Omega \\[4pt]
&\quad + \int_{\Omega} [s(\psi) R(C)\vartheta + Q_h] N_I N_J d\Omega \\[4pt]
V_{IJ} &= \int_{\Omega} (\boldsymbol{q} \cdot \nabla N_I) N_J d\Omega \\[4pt]
K_{IJ} &= \int_{\Omega} N_I (\boldsymbol{q} \cdot \nabla N_J) d\Omega \\[4pt]
G_{IJ} &= \int_{\Omega} \frac{\Delta t_n}{\varepsilon s(\psi)} (\boldsymbol{q} \cdot \nabla N_I) \cdot (\boldsymbol{q} \cdot \nabla N_J) d\Omega \\[4pt]
R_I &= \int_{\Omega} N_I s(\psi) Q_C d\Omega - \int_{\Gamma} N_I q_c d\Gamma
\end{aligned}
\right\} \qquad (13\text{-}63)
$$

REMARK 2. The PGLS damping matrix $\boldsymbol{G}$ in (13-63) is strongly related to the pore velocity $\boldsymbol{q}/[\varepsilon s(\psi)]$. For dry unsaturated problems with $s(\psi) \to 0$ the matrix $\boldsymbol{G}$ becomes singular.

REMARK 3. The used time marching schemes are second order accurate in time for $\theta = 1/2$ (Crank Nicolson scheme) and $\sigma = 2$ (Adams-Bashforth AB/TR predictor-corrector). They have a first order accuracy for $\theta = 1$ (fully implicit scheme) and $\sigma = 1$ (forward Euler/backward Euler FE/BE predictor-corrector).

REMARK 4. The PGLS symmetrization is caused by the fact that the advective matrices $\boldsymbol{V}$ and $\boldsymbol{K}$ form a symmetric contribution as the sum $(\boldsymbol{V} + \boldsymbol{K})$ because $\boldsymbol{K} = \boldsymbol{V}^T$ is the transpose. This is only attainable for a transport equation which has been written in the so-called convective form[1]. Divergence forms of transport equations cannot by handled by the PGLS scheme as derived in Section 13.2.5.

REMARK 5. The right-hand side of the matrix system (13-60) or (13-61) requires the assembly of the advective matrix $\boldsymbol{V}$ which is unsymmetric. As the consequence, quadrature rules common in the FEM to build up the element matrices have to be performed over all rows and columns of the submatrices, at least for the submatrix $\boldsymbol{V}$.

REMARK 6. A comparison of the streamline upwind Petrov-Galerkin (SUPG) scheme with the PGLS technique leads to the following: The 'balancing tensor diffusivity' of the SUPG is expressed by the streamline upwind term

$$\int_{\Omega} \nabla N_I \cdot \left(\beta^{\mathrm{num}} \frac{\boldsymbol{q} \otimes \boldsymbol{q}}{\|\boldsymbol{q}\|} \cdot \nabla N_J \right) d\Omega \qquad (13\text{-}64)$$

where β^{num} is the numerical (longitudinal) dispersivity as a free parameter which can be estimated as $\Delta l/2$ and $\Delta l/4$ for linear and quadratic finite elements, respectively. If comparing (13-64) of the SUPG with the PGLS damping matrix $\boldsymbol{G}$ in (13-63) it is obvious that both forms of the upwinding terms become identical if we set

---

1. see Chapter 6: *About the difference between the convective form and the divergence form of the transport equation,* p. 121 and following.

$$\frac{\beta^{num}}{\|q\|} \approx \frac{\Delta t_n}{\varepsilon s(\psi)} \qquad (13\text{-}65)$$

It reveals the upwind characteristic of the PGLS is quite similar to a streamline upwind technique. We note an equivalent damping parameter β^{num} for the PGLS becomes dependent on the time step size and the measure of the pore velocity according to $\Delta t_n \|v\|$.

land's results[18] for the case 1 are presented in Fig 13.5. The results of the present PGLS scheme compared with the standard Galerkin-FEM and the Crank-Nicolson time stepping scheme are displayed in Fig. 13.6 for the case 1. As seen in Figs. 13.5 and 13.6 there are no remarkable differences between the different schemes

13.4 Benchmarks

13.4.1 Two-dimensional advective-dominant transport at a grid-parallel flow

This example has been introduced by Wendland[18] to compare the symmetric streamline stabilization (termed as S3-scheme) to analytical results, to the operator split technique proposed by König[13] (termed as OS-scheme) and to the symmetric scheme developed by Leismann and Frind[14] (termed as L-scheme), which handles the advective term explicitly. The problem is described in Fig. 13.4.

Introducing the grid Peclet number Pg as

$$Pg = \frac{\|v\|\Delta l}{D} \qquad (13\text{-}66)$$

and the Courant number Co as

$$Co = \frac{\|v\|\Delta t}{\Delta l} \qquad (13\text{-}67)$$

we study three cases as listed in Table 13.1. Wend-

and in comparison with the exact (analytical) solution.
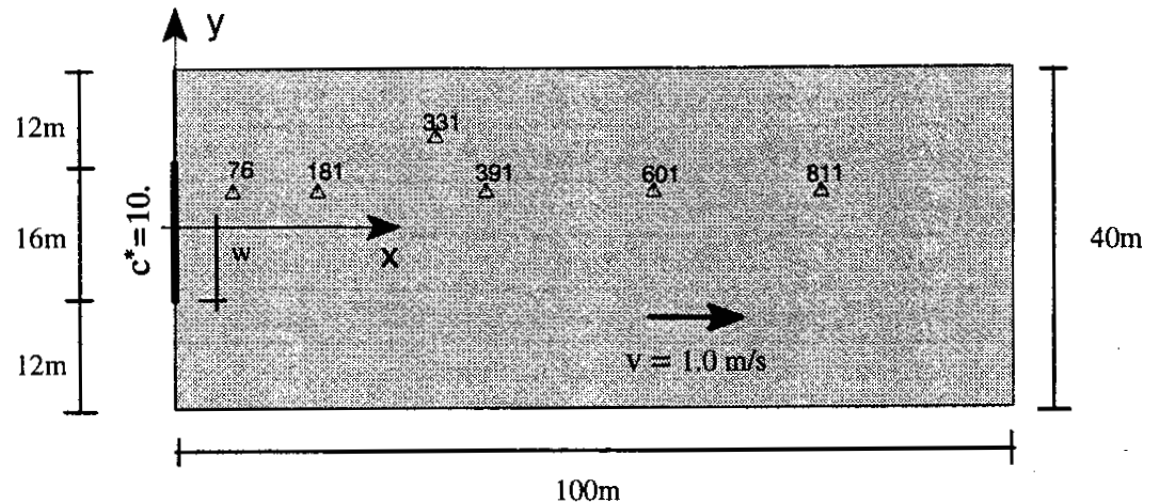
Figure 13.4 The 2D advection-dispersion problem with grid-parallel flow (from[18]).

Table 13.1 Simulation parameter used 2D advection-dispersion problem at grid-parallel flow

Case	Δt [s]	β_L [m]	β_T [m]	Δx [m]	Δy [m]	Pg	Co
1	1.0	2.0	0.2	2.0	2.0	1.0	0.5
2	2.0	0.02	0.002	2.0	2.0	100.0	1.0
3	2.0	0.02	0.002	0.5	0.5	25.0	4.0

Figure 13.6 Distribution of concentration at 50 s computed by a) the PGLS and b) the standard Galerkin-FEM for the case 1 with using the Crank-Nicolson time stepping scheme $\theta = 1/2$.

Figure 13.5 Distribution of concentration at 50 s computed by Wendland[18] for the case 1.

The case 2 represents an advective-dominant problem with a high grid-Peclet number of 100. Wendland[18] found the solutions as shown in Fig. 13.7. As seen its results are not satisfactory compared to the analytical solution. The L- and the OS-schemes reveal both oscillatory and overdiffusive solutions. Even false cross-damping effects are apparent. Wendland's S3-solutions improve the situation somewhat (see Fig. 13.7), however, wiggles and numerical dispersion effects, obviously due to the fully implicit approximation, appear.

The present PGLS strategy obtains better solutions which are exhibited in Fig. 13.8. The PGLS with the Crank-Nicolson scheme (Fig. 13.8a) echoes the best-

sharpest solutions compared to the implicit PGLS (Fig. 13.8b) and the standard SUPG with Crank-Nicolson (Fig. 13.8c). However, the concentration distribution for the Crank-Nicolson PGLS is not completely wiggle-free (Fig. 13.8a). On the positive side, unlike the L- and the OS-strategies the present schemes do not suffer from spurious cross-dispersion effects.
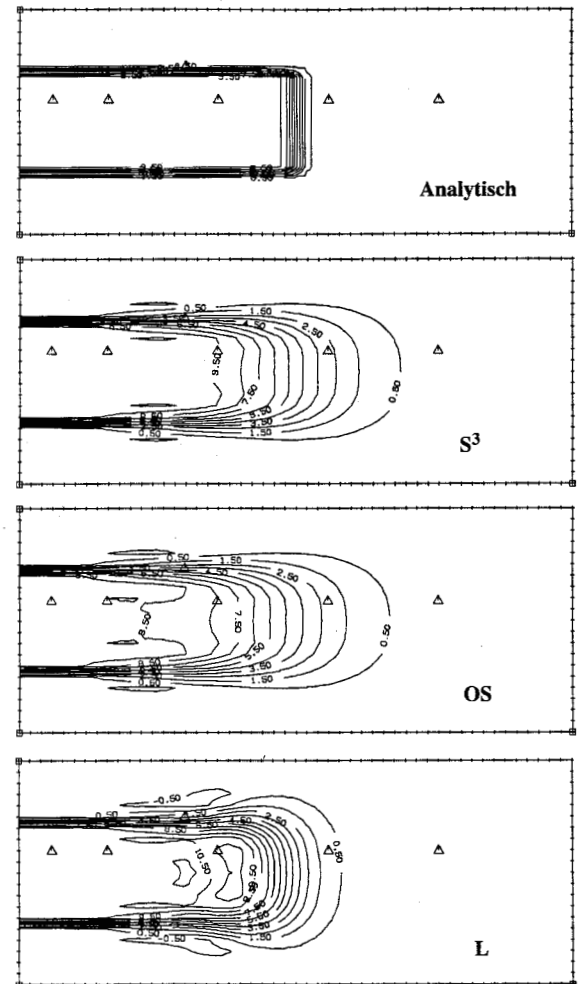
Figure 13.7 Distribution of concentration at 50 s computed by Wendland[18] for the case 2.
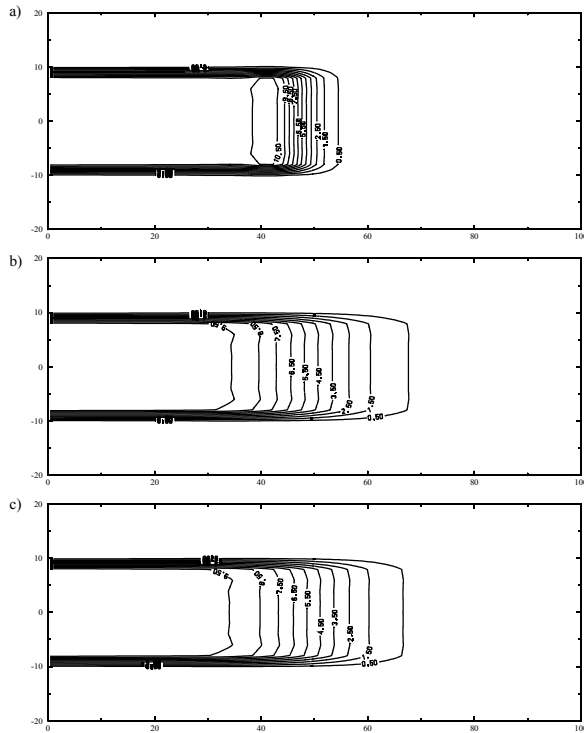
Figure 13.8 Distribution of concentration at 50 s com-
puted by a) the PGLS with the Crank-Nicolson scheme,
b) PGLS scheme with the fully implicit time marching,
and c) the standard streamline upwind (SUPG) with the
Crank-Nicolson scheme for the case 2.

The case 3 covers solutions of a refined grid. Wend-
land's results[18] are shown in Fig. 13.9. A certain reduc-
tion of the influence of numerical dispersion could be
achieved (*cf.* Figs. 13.9 and 13.5). The results of the
present PGLS technique are displayed in Fig. 13.10 for

the case 3. For the Crank-Nicolson PGLS (Fig. 13.10a)
wiggles in the concentration distribution are revealed.
On the other hand, a fully implicit PGLS (Fig. 13.10b)
becomes free of oscillations, but, smearing of the con-
centration indicates false numerical dispersion which is
mainly caused by the implicit technique of first order
accuracy in time. Unlike this, the standard SUPG tech-
nique with a Crank-Nicolson obtains reasonable results
without any wiggles and an acceptable amount of
damping (Fig. 13.10c).

Figure 13.10 Distribution of concentration at 50 s computed by a) the PGLS with the Crank-Nicolson scheme, b) PGLS scheme with the fully implicit time marching, and c) the standard streamline upwind (SUPG) with the Crank-Nicolson scheme for the case 3.

Figure 13.9 Distribution of concentration at 50 s computed by Wendland[18] for the case 3.

13.4.2 Two-dimensional advective-dominant transport at an oblique flow

To benchmark the PGLS technique for more general (non-parallel) flow conditions we consider the example which was also studied by Wendland[18]. The problem is described in Fig 13.11. With the flow boundary conditions an oblique flow in induced in the domain. The conductivity is assumed to $K = 10^2$ m/s. Two different cases are considered as listed in Table 13.1.
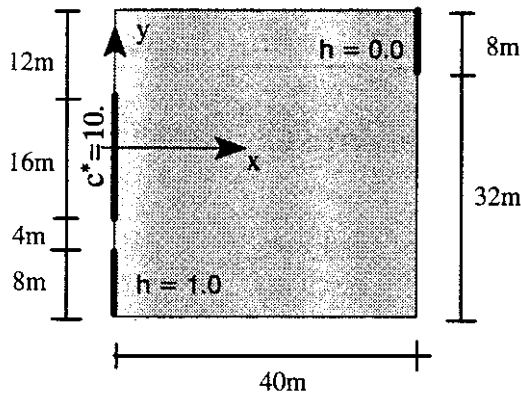
Figure 13.11 The 2D advection-dispersion problem with oblique flow (from[18]).

Wendland's results[18] for the case 4 are shown in Fig. 13.12, where the standard Ritz-Galerkin method (termed as RG), the symmetric streamline stabilization (termed as S3-scheme), the operator split technique (termed as OS-scheme) and the symmetric scheme developed by Leismann and Frind[14] (termed as L-scheme) are presented. As seen in Fig. 13.12 Wend-

land's results are rather mixed. Compared to the more accurate RG scheme its S3 solutions reveal an increased amount of numerical dispersion. The OS and L schemes are also depart from the RG results. We compare these results with our findings as shown in Fig. 13.13. There, we display three PGLS applications: one is the Crank-Nicolson PGLS using 20 time steps (Fig. 13.13a), the other is a fully implicit PGLS realization (Fig. 13.13b) and the third is the use of the adaptive AB/TR predictor-corrector PGLS (Fig. 13.13c) with 25 varying time steps. Since there is no analytical solution for the present problem we run the standard Galerkin-FEM with the AB/TR scheme on a dense mesh (25,921 nodes) as depicted in Fig. 13.13d. The agreement of the present PGLS results is quite well and provides better solutions as such found by Wendland[18] (*cf*. Fig. 13.12 and Fig. 13.13a-c with Fig. 13.13d).

Table 13.2 Simulation parameter used 2D advection-dispersion problem at oblique flow

Case	Δt [s]	β_L [m]	β_T [m]	Δx [m]	Δy [m]	Pg	Co
4	2.0	2.0	0.2	2.0	2.0	1.0	1.5
5	1.0	0.02	0.002	2.0	2.0	100.0	0.75
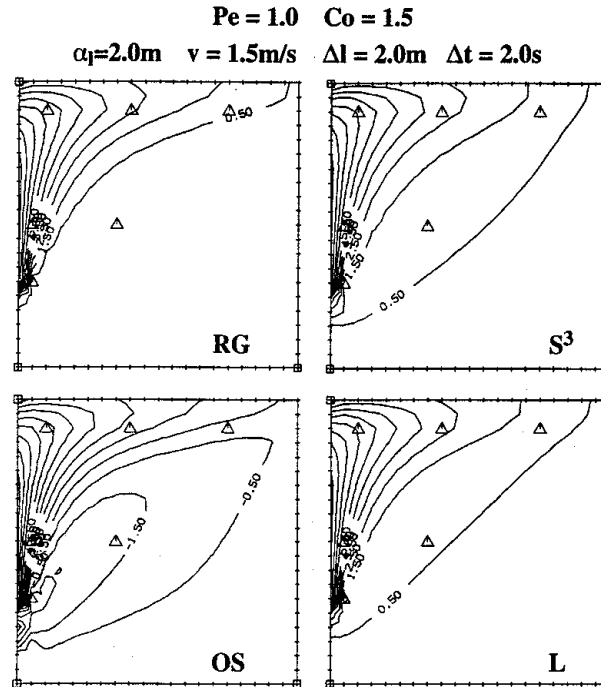
Figure 13.12 Distribution of concentration at 40 s computed by Wendland[18] for the case 4.
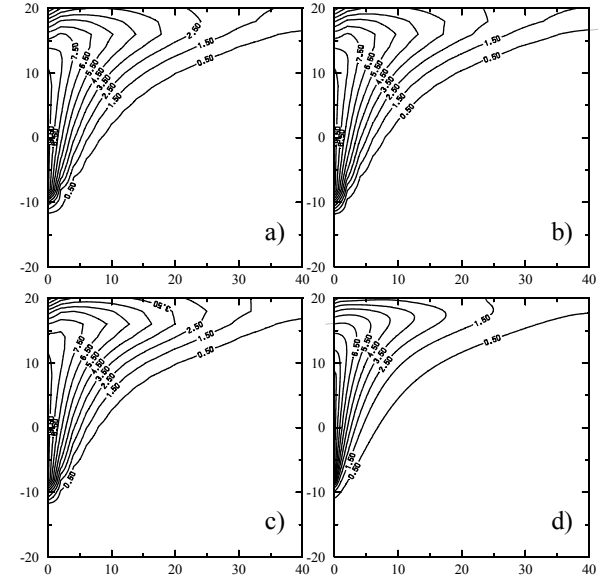
Figure 13.13 Distribution of concentration at 40 s computed by a) the Crank-Nicolson PGLS, b) fully implicit PGLS, c) AB/TR PGLS (25 adaptive predictor-corrector steps, RMS error = 10^{-3}) and d) the standard Galerkin-FEM with AB/TR predictor-corrector scheme applied to a dense mesh consisting of 25,921 nodes for the case 4.

Pe = 100.0   Co = 0.75

α_l=0.02m   v = 1.5m/s   Δl = 2.0m   Δt = 1.0s
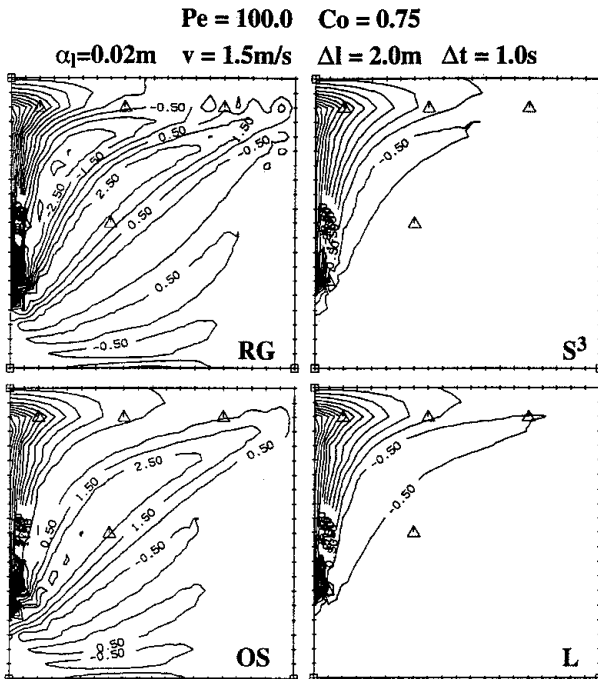
Figure 13.14 Distribution of concentration at 40 s computed by Wendland[18] for the case 5.
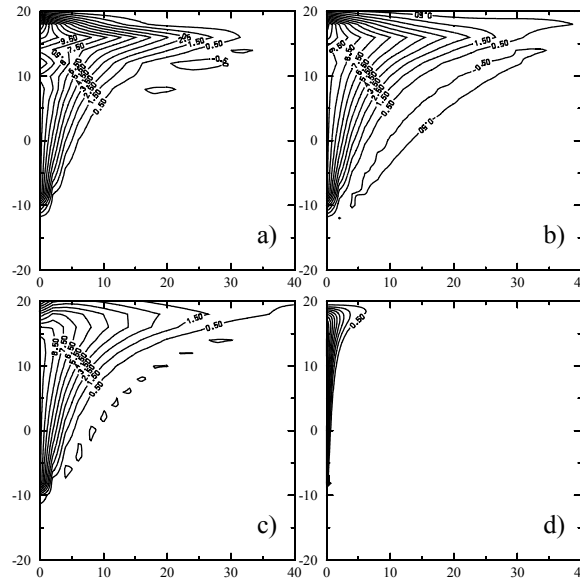
Figure 13.15 Distribution of concentration at 40 s computed by a) the Crank-Nicolson PGLS, b) fully implicit PGLS, c) Crank-Nicolson SUPG and d) the standard Galerkin-FEM with FE/BE predictor-corrector scheme applied to a dense mesh consisting of 25,921 nodes for the case 5.

The case 5 possesses very strong advection at a large grid Peclet number of $Pg = 100$. According to the formulation of the boundary-value problem the intruding amount of mass at the boundary is a function of the transverse dispersivity β_L. If holding small a very sharp boundary layer will be formed. It is to be expected, at coarse meshes false numerical dispersion

increases artificially the boundary layer which leads to an unphysically increase of intruding mass. This can be clearly seen in the results shown in Figs. 13.14 and 13.15 for case 5. Most important, we recognize that the PGLS can stabilize the solution (Fig. 13.14 (S3) and Fig. 13.15a,b). However, the solutions are not wiggle-free. Even in combination with the fully implicit time marching scheme the oscillations cannot be completely damped out. The PGLS results are comparable to the SUPG scheme (Fig. 13.15c). But, if comparing to a solution obtained on a dense mesh, we see the significant difference between results computed by the upwinding techniques (PGLS, SUPG) and the more accurate solution for a very fine mesh (Fig. 13.15d).

The test case 5 gives some insight into the characteristics of the PGLS scheme. Inspecting the matrix G of eqn (13-63) we see a dependence of the stabilization term on the time step size Δt. We can expect if using smaller time steps the influence of the stabilization term G reduces (or vanishes). We check this interdependence for the case 5 if reducing the constant time step size Δt at the given grid Peclet number $Pg = 100$. The obtained results are shown in Fig. 13.16 for four Co numbers. We note the following behavior. If reducing Δt somewhat the solution even improves in form of smaller oscillations in the concentration pattern (cf. Fig. 13.16a and 13.16b). But, if reducing further the step size Δt the PGLS produces wiggles again. The smaller the time step the more oscillations are generated as shown in Figs. 13.16c and 13.16d. We found at a sufficiently small Δt the solution becomes completely instable (Fig. 13.16d).
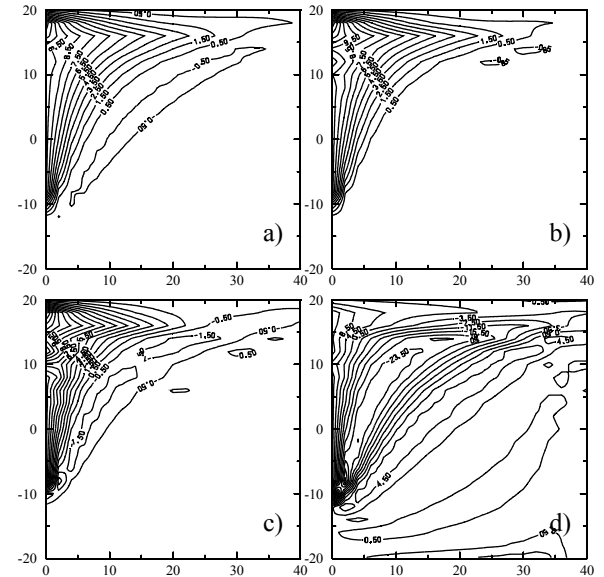
Figure 13.16 Distribution of concentration at 40 s computed by the fully implicit PGLS at different Courant numbers Co and given grid Peclet number of $Pg = 100$ (case 5): a) $Co = 0.75$ ($\Delta t = 1$s), b) $Co = 0.375$ ($\Delta t = 0.5$s), c) $Co = 0.12$ ($\Delta t = 0.16$s) and d) $Co = 0.075$ ($\Delta t = 0.1$s).

13.4.3 Three-and two-dimensional transport modeling of the patch source problem

The patch source problem refers to the transport of a solute from a boundary source of finite extent into a rectangular domain subjected to a uniform, unidirectional velocity field. It has been studied by Leismann and Frind[14] for 2D and Burnett and Frind[1,2] in 3D. The description of the benchmarks can be found in Segol's book[17]. The definition of the 3D problem is given in Fig. 13.17. The corresponding 2D problem refers to the cross-section in the *x-z*-plane. The parameters used for the 2D and 3D model are summarized in Table 13.3 and Table 13.4, respectively.
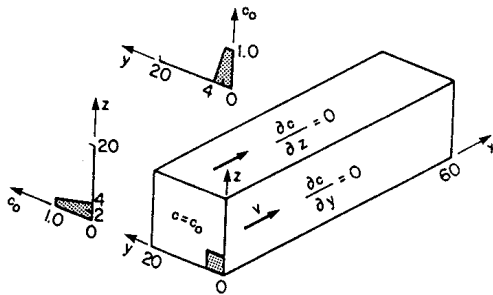
Figure 13.17 Definition of the patch-source problem in 3D (from[2,17]).

Table 13.3 Patch-source parameters used in 2D

Domain	Rectangle with dimension 200 x 40 in arbitrary units [L]
Source location	$x = 0, 0 \leq z \leq 8$
Pore velocity	0.1 [L/T] in the x-direction, 0.0 in the z-direction
Boundary conditions	$C_o = 1$ at the source, C = 0 outside the source at x = 0
Initial conditions	C = 0
Longitudinal dispersivity, β_L	1.0 [L]
Transverse dispersivity, β_T	1.0 [L]
Grid characteristics	Uniform grid with 2000 quadrilateral elements
Longitudinal spacing, Δx	2.0 [L]
Transverse spacing, Δz	2.0 [L]
Time increment, Δt	10 [T]
Grid Peclet number, Pg	2
Courant number, Co	0.5

Table 13.4 Patch-source parameters used in 3D

Domain	Parallelepiped of dimensions 60 x 20 x 20 in arbitrary units [L]
Source location	on the plane $x = 0$ in the region $0 \leq y \leq 3$, $0 \leq z \leq 3$
Pore velocity	0.1 [L/T] in the x-direction, 0.0 in the y- and z-directions
Boundary conditions	$C_o = 1$ at the source, $C = 0$ outside the source at $x = 0$ plane
Initial conditions	$C = 0$
Longitudinal dispersivity, β_L	1.0 [L]
Transverse dispersivity, β_T	0.25 [L]
Grid characteristics	Uniform grid with 30 x 20 x 20 hexahedral elements
Nodal spacing in x-direction, Δx	2.0 [L]
Nodal spacing in y-direction, Δy	1.0 [L]
Nodal spacing in z-direction, Δz	1.0 [L]
Time increment, Δt	20 [T]

The analytical solution for the 2D patch-source problem together with Leismann and Frind's symmetric matrix solution[14] are shown in Fig. 13.18. The present results obtained by the Crank-Nicolson PGLS scheme are plotted in Fig. 13.19. It becomes evident the agreement is quite perfect and a high accuracy could be achieved with the PGLS technique.
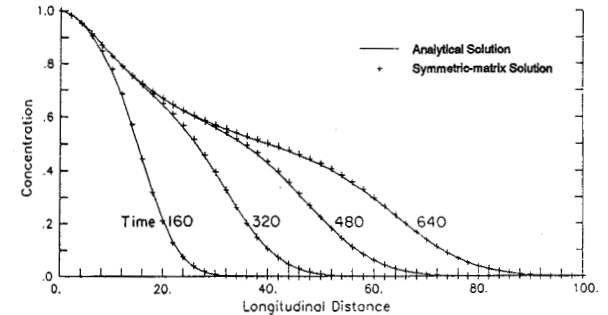
Figure 13.18 Concentration profiles along the center of the plume for the 2D patch-source problem - Analytical and Leismann and Frind's results[14].
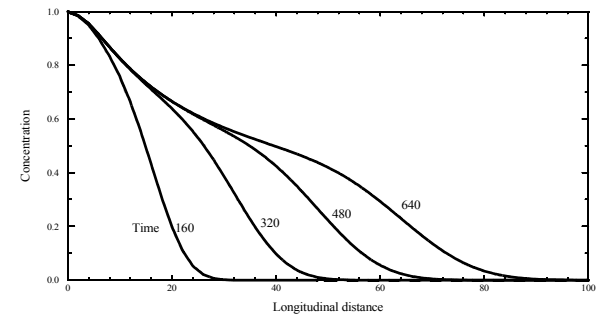
Figure 13.19 Concentration profiles along the center of the plume for the 2D patch-source problem - Crank-Nicolson PGLS results.

The results computed by Burnett and Frind[1,2] for the 3D patch-source problem are depicted in Figs. 13.21

and 13.23 for selected concentration profiles. The shown diagrams also involve analytical solutions which exist for the 3D patch-source problem. Burnett and Frind could achieve good agreements with the analytical results and our results obtained for both the Crank-Nicolson PGLS and the Crank-Nicolson standard Galerkin-FEM also lead to a satisfactory agreement as exhibited in Figs. 13.22 and 13.24 for representative concentration profiles. However, we note there are slight differences between the standard GFEM and the PGLS scheme, where the GFEM gives a better agreement with the analytical solution. The 3D view of the obtained PGLS results and the used finite element mesh are presented in Fig. 13.20.
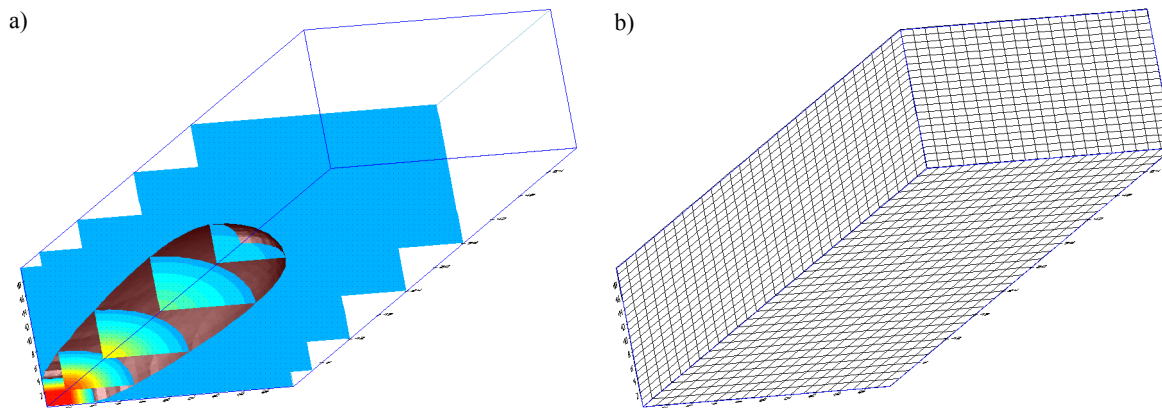
Figure 13.20 a) 3D view (from bottom to top) of the plume distribution obtained by the PGLS scheme and b) used finite element mesh.
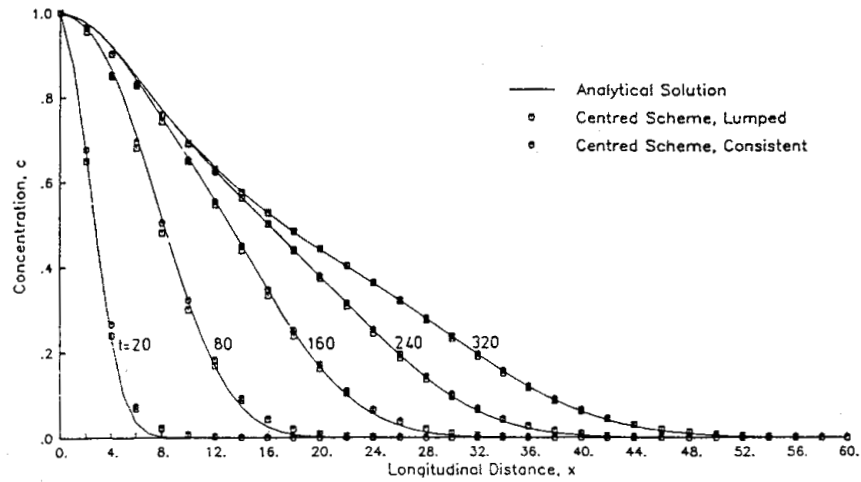
Figure 13.21 Longitudinal concentration profiles at $y = 0$ and $z = 0$ for the 3D patch-source problem - Analytical and Burnett and Frind's results [1,2,17].
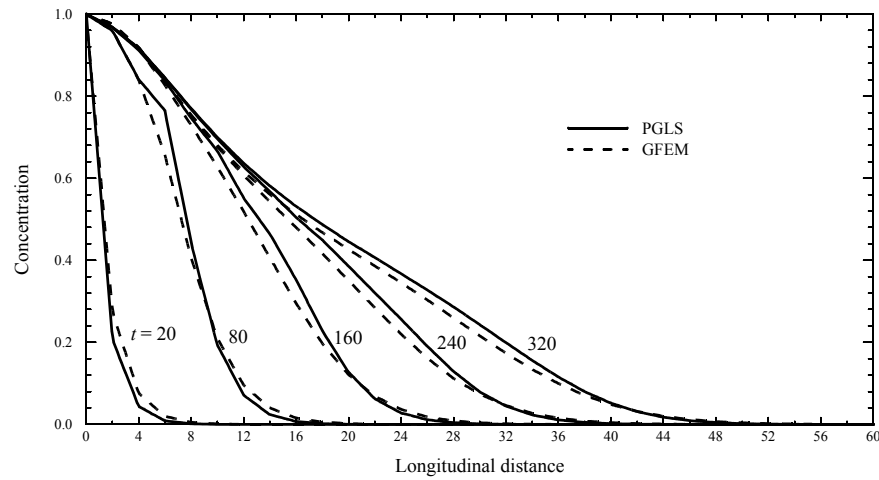
Figure 13.22 Longitudinal concentration profiles at $y = 0$ and $z = 0$ for the 3D patch-source problem - Crank-Nicolson PGLS and standard GFEM results.
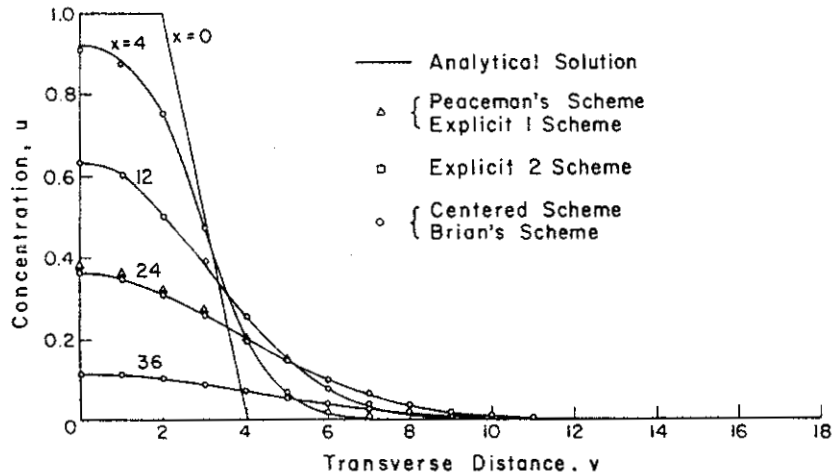
Figure 13.23 Transverse concentration profiles at $y = 0$ and $z = 0$ and at the time stage of 320 for the 3D patch-source problem - Analytical and Burnett and Frind's results[1,2,17].
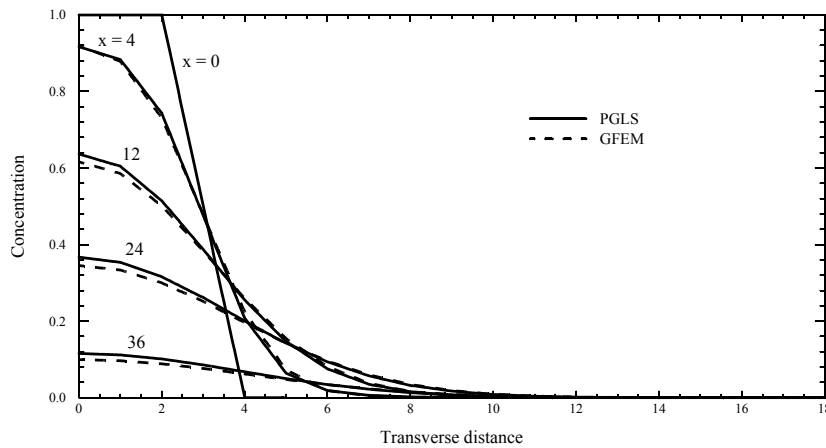
Figure 13.24 Transverse concentration profiles at $y = 0$ and $z = 0$ and at the time stage of 320 for the 3D patch-source problem - Crank-Nicolson PGLS and standard GFEM results.

13.4.4 Hoopes and Harlemann's two-well problem

Hoopes and Harlemann[8] performed a lab-scale experiment in a semi-cylinder filled with sand (Fig. 13.25). They measured the distribution of a solute between a recharge and a pumping well. For an analytical solution they set up a conceptual model of a two-dimensional horizontal confined aquifer which is homogeneous and isotropic. The flow between the well doublette at a distance, $2d = 0.61$ m, is isothermal and in a steady state. The solute transport is advective, dispersive along streamlines and has a molecular diffusion. Comparisons of the analytical result with experiments and various numerical solution schemes have already been performed elsewhere[8,11,17]. For the present benchmark calcula-tions we focus on the newly introduced PGLS scheme and check the results against the analytical results, the standard Galerkin (GFEM) and the streamline upwinding (SUPG) schemes of FEFLOW.

One obtains the analytical solution in terms of the velocity potential Φ and the streamline function Ψ. They are related to the original x,y-coordinates via the conformal transformation

$$\Phi + i\Psi = \text{Ln}(z+d)/(z-d) \qquad (13\text{-}68)$$
$$\text{with} \qquad z = x + iy$$

This transformation maps the area of the half circle with radius $r \le d$ onto a strip of infinite length and width $\pi/2$. The transport equation transforms to

$$\frac{\partial C}{\partial t} + v^2\left(\frac{\partial C}{\partial \Phi} + D\frac{\partial^2 C}{\partial \Phi^2}\right) = 0 \qquad (13\text{-}69)$$

and is now one-dimensional with the symbols taken from Appendix A. From our assumptions on the nature of the transport process we obtain the dispersion coefficient $D = \beta_L v + D_d$.
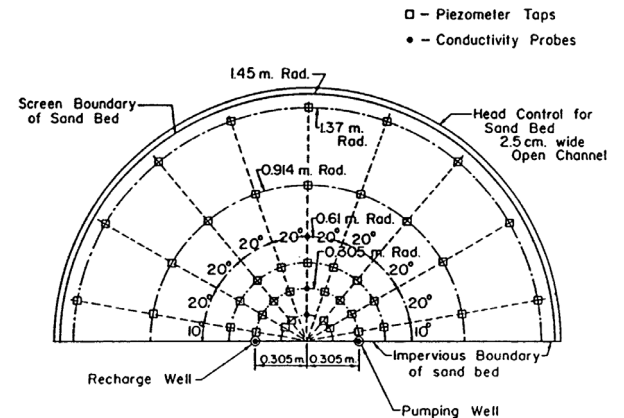
Figure 13.25 Plane view of Hoopes and Harlemann's sand-filled semi-cylinder.

The pore velocity v at a flux rate Q of the recharge well is

$$v = \frac{q}{\varepsilon} = \frac{Q}{2\pi M\varepsilon d}(\cosh(\Phi_D) + \cos(\Psi_D)) \qquad (13\text{-}70)$$

with the dimensionless quantities

$$\Phi_D = \frac{1}{2}\ln\frac{(x+d)^2+y^2}{(x-d)^2+y^2} \qquad \text{and} \qquad (13\text{-}71)$$

$$\Psi_D = \operatorname{atan}\frac{-2yd}{x^2+y^2-d^2}$$

The velocity potential $\Phi(x,y)$ and the streamline function $\Psi(x,y)$ are obtained after multiplication with $Q/2\pi M$. The initial and boundary conditions are

$$C(0,\Phi,\Psi) = 0 \qquad \text{and} \qquad (13\text{-}72)$$
$$C(t,\Phi(-d,0),\Psi(-d,0)) = C_0$$

The dimensionless concentration at arbitrary time is given by

$$C_D = \frac{C}{C_0} = \frac{1}{2}\operatorname{erfc}\left(\frac{I_D-t_D}{2\sqrt{J_D}}\right) \qquad (13\text{-}73)$$

with the dimensionless time $t_D = Q/(2\pi\varepsilon M d^2)\cdot t$. Owing to the properties of the conformal transformation (13-68) the concentration can only be calculated for spatial points with $r \le d$.

The integrals

$$I_D = \int_{-\infty}^{\Phi_D}\frac{d\Phi'_D}{v_D^2} \qquad \text{and} \qquad J_D = \int_{-\infty}^{\Phi_D}\frac{D_D}{v_D^4}d\Phi'_D \qquad (13\text{-}74)$$

with $v_D = \cosh(\Phi_D)+\cos(\Psi_D)$ and $D_D = 2\pi M\varepsilon D/Q$ are likewise dimensionless and have been integrated numerically. A complete analytical solution is possible but cumbersome.

We did the numerical FEM analysis on the mesh of Fig. 13.26 which has been refined in the half circle with $r \le d$ where high Darcy fluxes occur. This measure lowers the grid Peclet number there. For the time step control we employed the forward Euler/backward Euler (FE/BE) predictor-corrector method. Hoopes and Harlemann[8] assumed no dispersion across streamlines in their formulation of Eq. (13-69). The longitudinal dispersivity $\beta_L = 0.0015$ m is very small and gives rise to a steep concentration front. Hence, the transport is dominated by advection as shown in Fig. 13.27 for breakthrough curves at two observation points along the symmetry line between the wells. The analytical curve with the steepest slope can only be approximated by numerical schemes due to finite lattice element sizes. Furthermore, well-known problems of oscillating numerical solutions appear. The SUPG scheme has been employed to dampen the oscillations but introduces in turn additional non-physical dispersion. The GFEM scheme without upwinding techniques shows slight oscillations only for high concentrations at the second observation point $y = 0.305$ m which may be attributed to the coarser mesh in that region. The dispersion is reasonably low. Large wiggles occur for the PGLS scheme for high concentrations at both points. Even the onset of the breakthrough curves is not wiggle-free. For intermediate concentrations the PGLS front matches the analytical solution quite close.

three numerical schemes is depicted in Fig. 13.28. The analytical solution has been omitted here since it is only available in the half circle with $r \leq d$. The center of the concentration front at $C/C_0 = 0.5$ is the same in the whole region for all schemes. The SUPG scheme again exhibits the expected additional dispersion. Almost wiggle-free appears the GFEM scheme with an exception at some distance above the recharge well. The PCG solution oscillates considerably along the concentration fringe of $C/C_0 = 1$. Larger oscillations appear left of the region between the wells. But wiggles can also be observed in the direction of the pumping well. This asymmetry is again due to the finer mesh between the wells.

Table 13.5 Parameter of the FEFLOW for Hoopes and Harlemann's two-well problem

Name	Symbol	Value
Steady flow		
flow boundaries		
flux at recharge well	Q	$6.4327 \ \text{m}^2/\text{d}$
hydraulic head at pumping well	h	$0 \ \text{m}$
flow materials		
aquifer transmissivity	T	$1 \cdot 10^{-4} \ \text{m}^2/\text{s}$
Transient transport		
transport initials		
homogeneous concentration	$C(t=0, x, y)$	$0 \ \text{mg/l}$
transport boundaries		
concentration at recharge well	C_0	$1 \ \text{mg/l}$
transport materials		
aquifer thickness	M	$0.089 \ \text{m}$
porosity	ε	0.374
adsorption	χ	0
molecular diffusion	D_d	$0 \cdot 10^{-9} \ \text{m}^2/\text{s}$
longitudinal dispersivity	β_L	$0.0015 \ \text{m}$
transverse dispersivity	β_T	$0 \ \text{m}$
decay rate	ϑ	$0 \cdot 10^{-4} \ /\text{s}$
FEM		
wellbore radius	r_B	$0.05 \ \text{m}$
outer boundary radius	R	$1.45 \ \text{m}$
mesh: 3-noded triangle elements		
number of nodes	n_p	1554
number of elements	n_e	2950
time stepping regime: automatic time step control		
initial time step	Δt_0	$10^{-5} \ \text{d}$
simulation time period	T	$0.2 \ \text{d}$
solver: direct Gauss elimination		

Figure 13.26 Finite element grid used.

The concentration distribution at $t = 0.05$ d for all
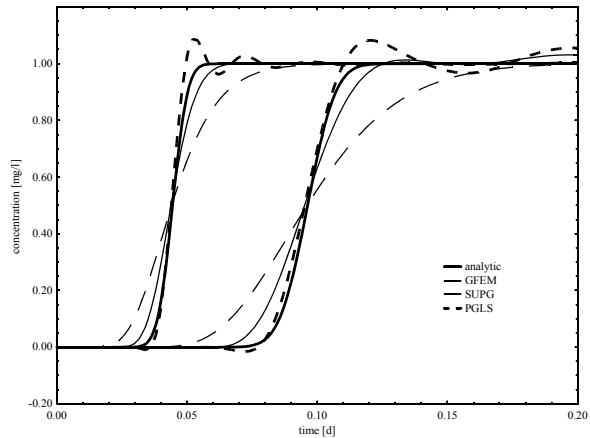
Figure 13.27 Breakthrough of the concentration at two points $y = 0.145$ m (left) and 0.305 m (right) on the symmetry line $x = 0$ between the wells.

In this benchmark calculation the GFEM scheme is the method of first choice with reasonable dispersion and very few oscillations. The PGLS scheme has not performed sufficiently well due to the appearance of wiggles with unacceptable amplitudes both for small and high concentrations.
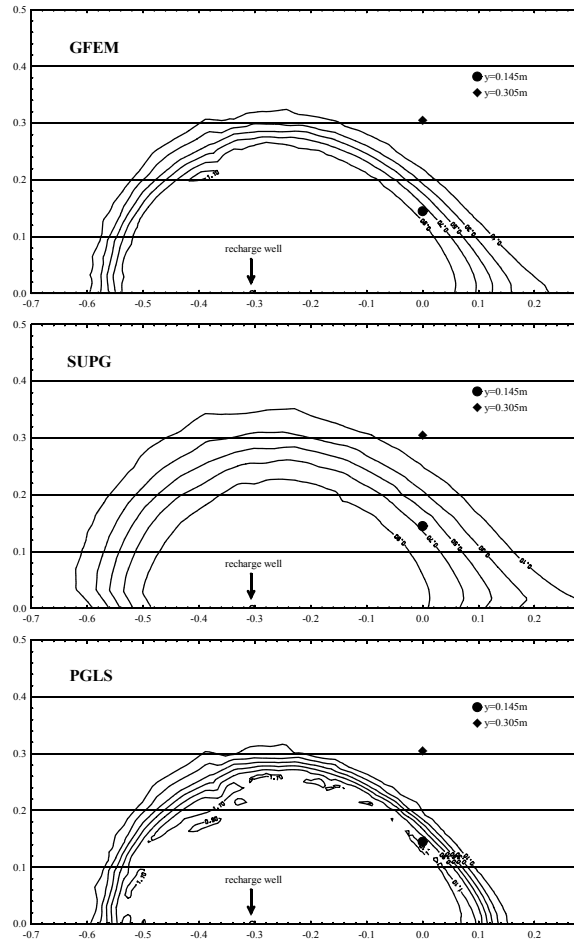
Figure 13.28 Distribution of concentration computed by GFEM, SUPG and PGLS at $t = 0.05$ d.

13.4.5 Highly advective solute transport in a steady-state

This benchmark is taken from Nguyen's technical note[16]. Nguyen employed a two-dimensional flow field of the form

$$q_x = \frac{1}{2}(3z - 1) \qquad \text{and} \qquad q_z = 0 \qquad \text{(13-75)}$$

to compare his PGLS scheme with the Taylor-Galerkin finite element splitting-up method of Donea *et al.*[7] in the case of advective-diffusive solute transport in a steady state. He found the PGLS scheme superior in situations where advection severely dominates the solute transport.

Nguyen[16] introduced the flow field (13-75) in a square cavity of unit edge length as shown in Fig. 13.29. He performed his calculations on a mesh with 20 x 20 bilinear elements. For the transport problem he studied the concentration distribution in the advection-dominated regime for two different grid Peclet numbers $Pg = |q_x^{max}| \Delta z / 2D$. He used $Pg = 1.25$ and $Pg = 5$ which lead to diffusivities of 0.02 and 0.005, respectively, in arbitrary units. The initial concentration and the concentration flow across the horizontal boundaries were zero. A concentration gradient of unity is given by the right and left boundary concentrations.

The FEFLOW calculations were done for a two-dimensional horizontal confined aquifer of thickness 1 m. Setting up the simulation model within FEFLOW is not straightforward since the curl of the flow field (13-75) does not vanish, i.e. $\nabla \times q \neq 0$. Such fields do not occur for groundwater flow problems. However, the flow direction can be forced almost entirely into the x-direction by making the aquifer very anisotropic with a factor of 10^{-5}. We did the calculations for steady-state flow and transient transport with the PGLS scheme. All FEFLOW parameters for this benchmark are listed in Table 13.6. Their values can be derived from Nguyen's original problem if one sets the dimensionless edge length to 1 m and introduces a concentration unit of 1 mg/l.
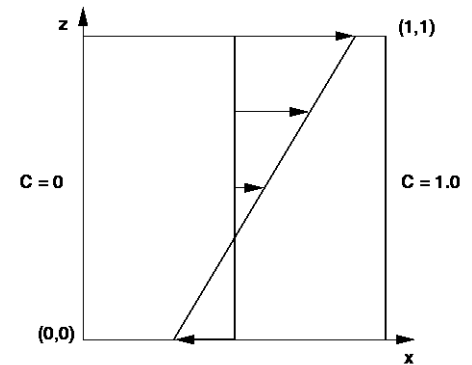
Figure 13.29 Square cavity with flow field and concentration boundary conditions.

The comparison of the concentration distribution at steady-state for $Pg = 1.25$ (Fig. 13.30) and $Pg = 5$ (Fig. 13.31) between Nguyen's[16] and the present FEFLOW calculations shows almost identical results in both cases. However, we found it very difficult to choose appropriate time stepping regimes. For the smaller Peclet number we used constant time steps at $\Delta t = 0.01$ d which means a maximal Courant number $Co = 0.2$ for a grid element. For the larger Peclet number we took a time step of $\Delta t = 0.035$ ($Co = 0.7$). In this case of high advection the PGLS results react sensitively to a

change of the Courant number. Lower Courant numbers produce wiggles in the critical regions next to the horizontal boundaries. Higher Courant numbers (or time steps) render concentration distributions with too much dispersion.

Obviously there exists only an interval of time step lengths where the PGLS scheme remains stable. We have already observed a similar behavior in the test case 5 of benchmark 3.2. With this test case we found that the main reason lies in the time step dependence of the stabilization matrix G in Eq. (13-63).

To complete our analysis of this benchmark we show the FEFLOW concentration distribution of a steady-state flow and transport calculation with a dense mesh of 6561 nodes and 12800 3-noded-triangle elements for the dimensionless diffusivity $D = 0.005$ in Fig. 13.32. The grid Peclet number is now reduced from $Pg = 5$ to $Pg \approx 1$. Here the dispersion at the upper right and the lower left boundaries both of Nguyen's and the FEFLOW results has been removed by a high lattice resolution.

Table 13.6 Parameter for highly advective solute transport in a steady-state

Name	Symbol	Value
Steady flow		
flow boundaries		
flux at right/left boundary	q_x	$\pm\frac{1}{2}(3z-1)$ m^2/d
flow materials		
aquifer transmissivity	T	$1 \cdot 10^{-4}$ m^2/s
aquifer anisotropy		10^{-5}
Transient transport		
transport initials		
homogeneous concentration	$C(t=0,x,z)$	0 mg/l
transport boundaries		
concentration along right/left boundary	$C_{r,l}$	1 mg/l; 0 mg/l
transport materials		
aquifer thickness	M	1 m
porosity	ε	1
adsorption	χ	0
molecular diffusion for $Pg = 1.25$	D_d	$231.482 \cdot 10^{-9}$ m^2/s
molecular diffusion for $Pg = 5$	D_d	$57.8704 \cdot 10^{-9}$ m^2/s
longitudinal dispersivity	β_L	0 m
transverse dispersivity	β_T	0 m
decay rate	ϑ	0 /s
FEM		
edge length of square	l	1 m
mesh: 4-noded quadrilateral elements		
number of nodes	n_p	441
number of elements	n_e	400
time stepping regime: constant time steps		
$\Delta t = 0.01$ d for $Pg = 1.25$; $\Delta t = 0.035$ d for $Pg = 5$		
simulation time period	T	10 d
solver: direct Gauss elimination		
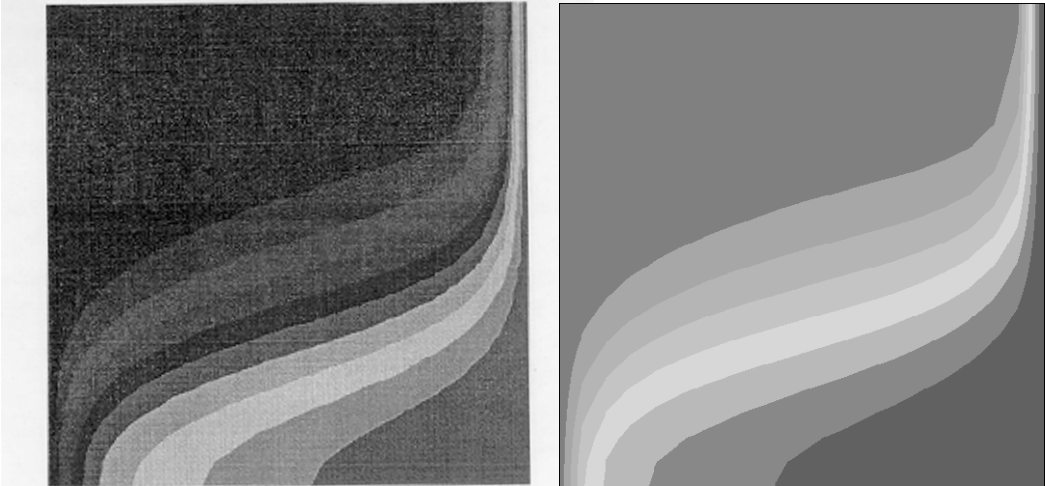
Figure 13.30 Concentration distribution at steady-state and $Pg = 1.25$, Nguyen[16] (left) and FEFLOW (right).
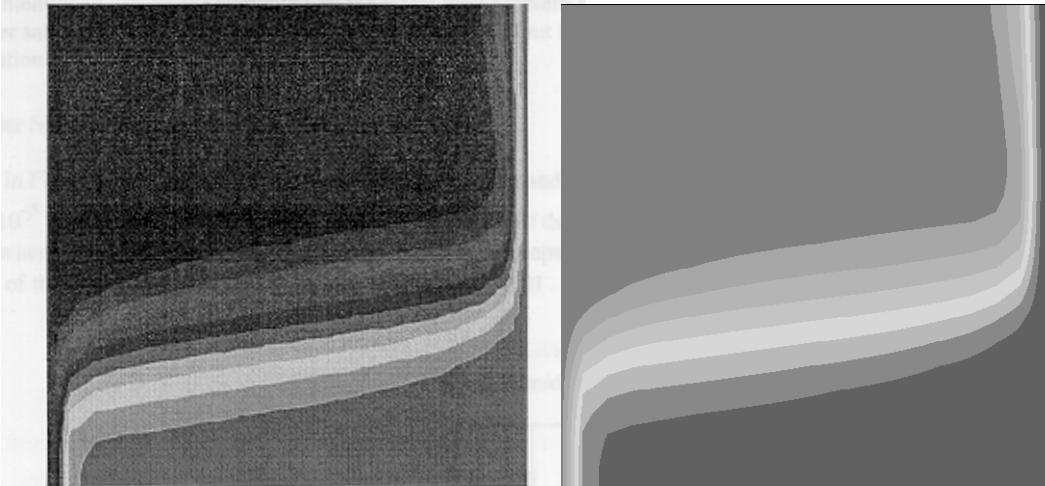
Figure 13.31 Concentration distribution at steady-state and $Pg = 5$, Nguyen[16] (left) and FEFLOW (right).
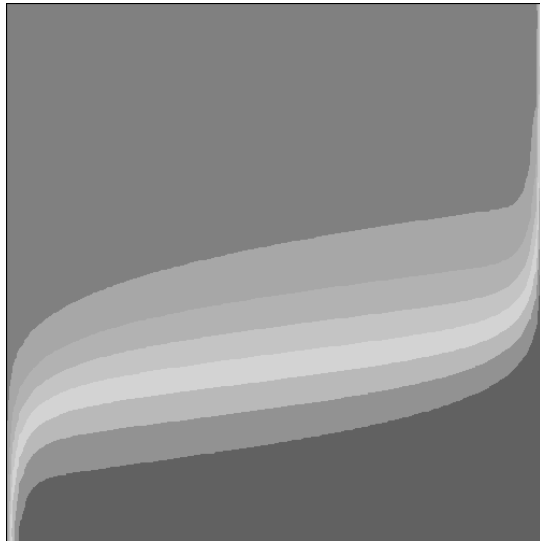
Figure 13.32 Concentration distribution at steady-state and $D = 0.005$ for a dense mesh.

1 m in the top half of the slab. Then it increases linearly to -0.01 m at the bottom. Since the hydraulic head h is related to the pressure head ψ by adding the elevation z, namely $h = \psi + z$, the initial flow will be completely downward from the top half to the bottom half even if the latter is more saturated. We expect the top half of the column to fall dry except at the right where a top inflow provides enough water to prevent this process.
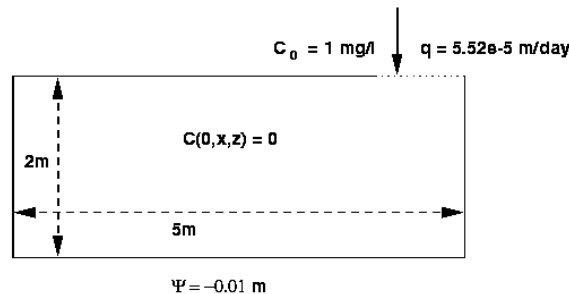
Figure 13.33 Problem measure, flow and transport conditions for a vertical rectangular slab of Kent soil .

13.4.6 Two-dimensional unsaturated flow and transport

Nguyen[16] has set up a flow and transport model of an unsaturated soil slab as shown in Fig. 13.33. He considered a vertical rectangle of 5 m width and 2 m depth. Then he applied a small Darcy flux of $5.52 \cdot 10^{-5}$ m/d across 1 m at the top in the right edge of the slab. The bottom of the slab is held at a small negative hydraulic pressure ψ of -0.01 m which corresponds to a high water content of a capillary fringe in the vicinity of a water table. The initial vertical pressure head is set to -

Nguyen used a soil moisture retention curve

$$s = \frac{a}{a + b|\psi|^c} \tag{13-76}$$

and a relative conductivity relationship $K_r(s)$

$$K_r = \left(\frac{s - s_r}{1 - s_r}\right)^d \tag{13-77}$$

with parameters which are typical for a very impermeable clay of the North Kent marshes in England. The

symbols are explained in Appendix A and the parameters are listed in Table 13.7.

Table 13.7 Parameters for unsaturated flow in Kent soil from Nguyen[16]

Name	Symbol	Value
saturated conductivity	K_s	$5 \cdot 10^{-9}$ m/s
porosity	ε	0.42
residual saturation	s_r	0.02
	a	5.30
	b	2.10 m^{-c}
	c	0.39
	d	18.8

The relationships (13-76) and (13-77) do not match both together any of FEFLOW's parametric models for unsaturated flow. Only the soil moisture curve (13-76) corresponds to Haverkamp's relation

$$s = s_r + \frac{\alpha(s_s - s_r)}{\alpha + |\psi|^\beta} \qquad (13\text{-}78)$$

if we choose the parameters appropriately. The choice is straightforward with the parameters listed in Table 13.8. For the relative conductivity we are now forced to use

$$K_r = \frac{A}{A + |\psi|^B} \qquad (13\text{-}79)$$

from the Haverkamp model. Thus, we must fit

Haverkamp's parameters A, B to match Nguyen's relation (13-77). The fitting process is difficult because satisfactory results for the whole region of K_r which stretches over many orders of magnitude cannot be obtained. This is due to the nature of the relation (13-77) which decays algebraically with no inherent scaling conductivity. We therefore choose to fit the relation for a soil water saturation $s \geq 0.6$ because in the soil slab relatively wet conditions prevail. Figure 13.34 shows the comparison of the fitted curve and Nguyen's original curve. The fitted curve matches sufficiently well for $s \geq 0.6$ but cannot be applied in dry conditions. The resulting values for Haverkamp's parameters A, B are listed in Table 13.7.

Table 13.8 Parameters for unsaturated flow in Kent soil in the FEFLOW model

Name	Symbol	Value
saturated conductivity	K_s	$5 \cdot 10^{-9}$ m/s
porosity	ε	0.42
residual saturation	s_r	0.0
maximum saturation	s_s	1.0
	$\alpha = a/b$	2.5238
	$\beta = c$	0.39
	A	0.01744
	B	0.6548
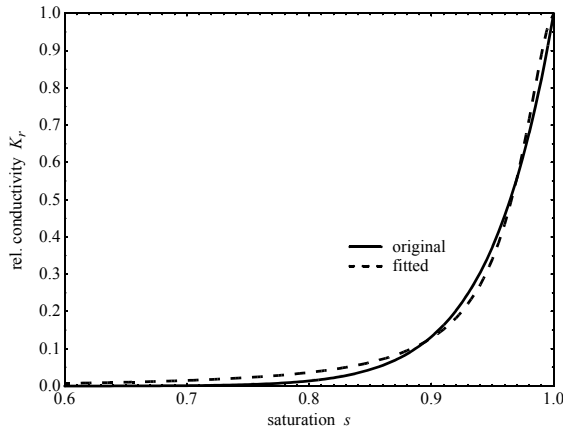
Figure 13.34 Relative conductivity relation for Kent soil as used in Nguyen[16] and in the present FEFLOW model.

The numerical calculations have been carried out on a 50x20 mesh of bilinear square elements. We employed the forward Euler/backward Euler predictor-corrector time stepping technique (FE/BE) for the flow and transport calculation. Figure 13.35 compares the results of Nguyen and the present calculation after $t = 400$ d for the pressure head ψ. They agree very well despite the slightly different relationships for the relative conductivity.

At $t = 400$ d the wetting front has not yet reached the bottom of the slab. This is the case in Fig. 13.36 after 10,000 days but the steady state is not yet reached. In this state the whole region is (almost) saturated since water is continuously infiltrated at the top but cannot leave the slab via the bottom boundary.

For the transport problem one can do a rough esti-

mation of the propagation of the solute front by multiplying the top boundary flux which determines the velocity in the system by the simulation time. One can divide this number by the pore velocity and obtains a displacement of ca. 1 m after 10,000 days. Around this time a concentration front should be well visible in the slab.

For the FEFLOW computation we used the transport parameter of Table 13.9. A boundary concentration of 1 mg/l has been imposed along the boundary flux line at the top as shown in Fig. 13.33. The parameters correspond to those of Nguyen except that we did not allow for decay and adsorption. Therefore our concentration contours should have a larger extent than those of Nguyen at a given time step. Unfortunately, we were unable to reproduce this behavior for the transport calculation. According to our estimation from above a sizeable concentration distribution should appear at some thousand days. This in contrast to Nguyen's findings whose results suggest that the front reaches the half depth after some hundred days. We do not show his results here but this discrepancy of an order of magnitude in the time of equal solute displacement still needs to be clarified.

We now proceed to the comparison of the GFEM, SUPG and PGLS schemes of FEFLOW for the concentration distributions at 5000 days (Fig. 13.37) and 10,000 days (Fig. 13.38). First we state that our estimation made a reasonable prediction for the extent of the concentration front. Before 5000 days all three schemes render wiggle-free solutions. From then on wiggles appear with the GFEM scheme but the PGLS scheme remains stable.
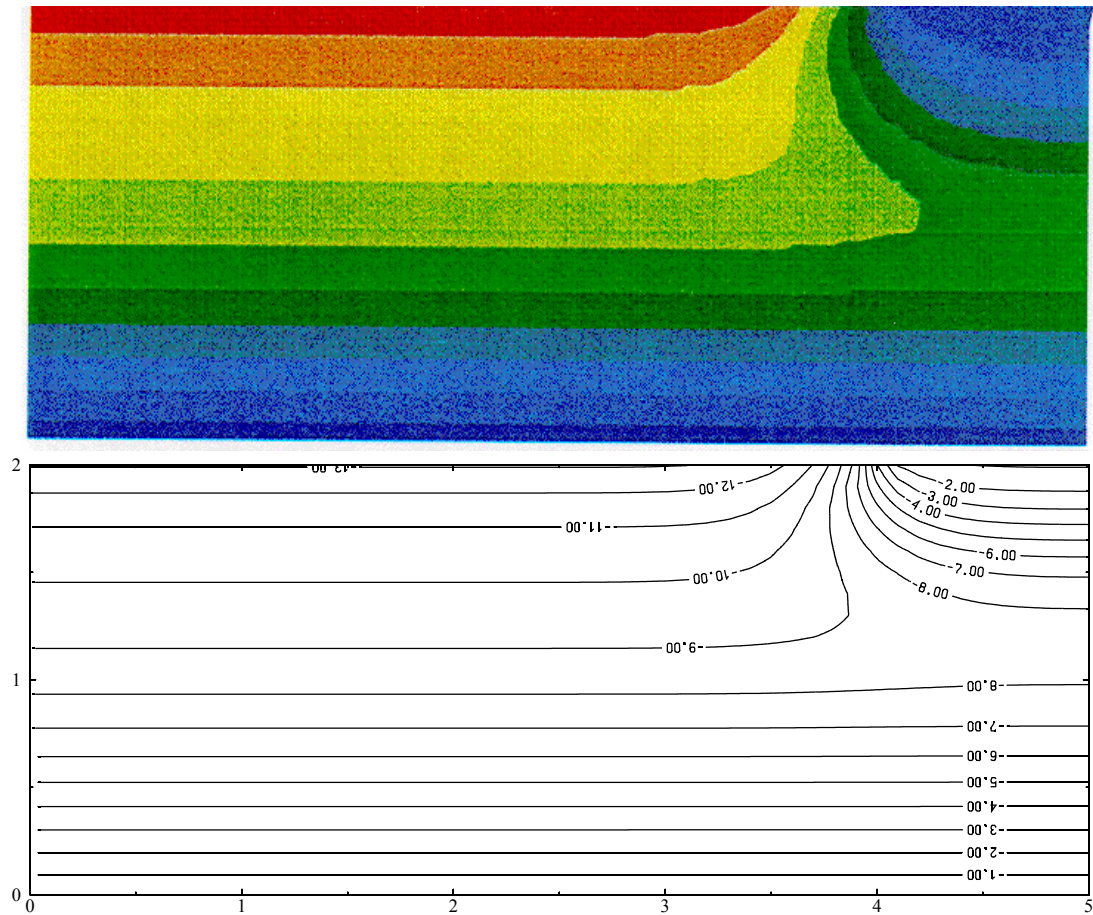
Figure 13.35 Pressure head ψ distribution after 400 days from Nguyen[16] (top) and the present FEFLOW computation (bottom), contour line units are [kPa].

At 10,000 days the PGLS scheme creates smaller oscillations as well but for the GFEM scheme they extend over the whole concentration front. The SUPG scheme remains wiggle-free as expected but smoothens the concentration front in an unphysical way. The wiggles are not caused by the mesh which is homogeneous in the whole region. They emanate from the boundary point with the coordinates (4 m, 2 m) where the bound-

ary flux jumps from a finite value to zero. High flux gradients occur in the vicinity of this point and are most difficult to reproduce by the GFEM scheme whereas the PGLS scheme is more successful.

We therefore prefer the PGLS scheme in this bench-

mark calculation. It is noteworthy, though, that even for the moderate conditions of a slow flow in a relatively wet soil wiggles still cannot be suppressed by this scheme.
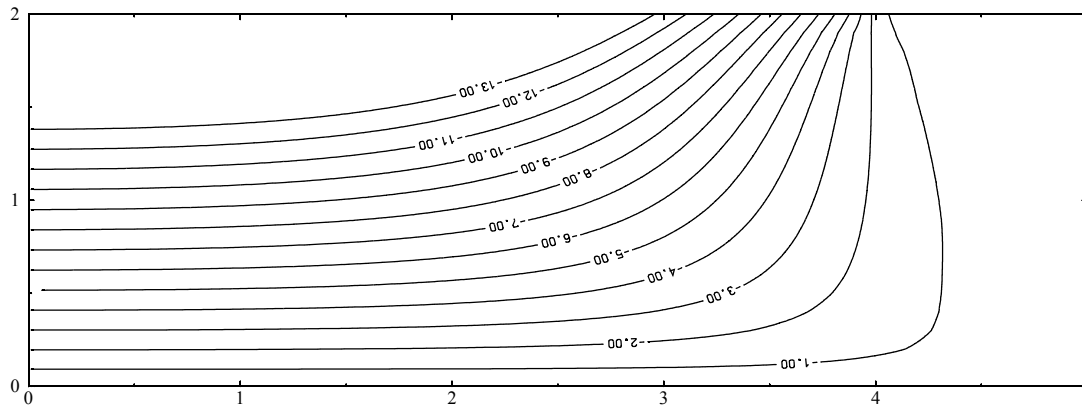
Figure 13.36 Pressure head ψ distribution after 10,000 days for the present FEFLOW computation, contour line units are [kPa].

Table 13.9 Transport parameters for Kent soil in the FEFLOW model

Name	Symbol	Value
longitudinal dispersivity	β_L	10^{-2} m
transverse dispersivity	β_T	0 m
molecular diffusion	D_d	$1.1574 \cdot 10^{-11}$ m^2/d
porosity	ε	0.42
decay rate	ϑ	0.0
adsorption	χ	0.0
boundary condition	C_0	1 mg/l
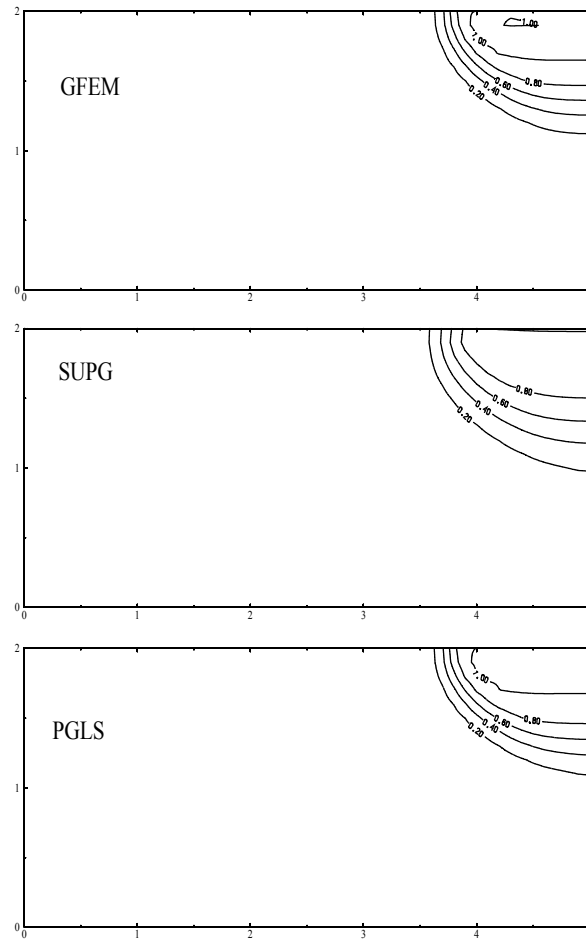
Figure 13.37 Concentration distribution after 5000 days for the GFEM, SUPG and PGLS schemes of FEFLOW.
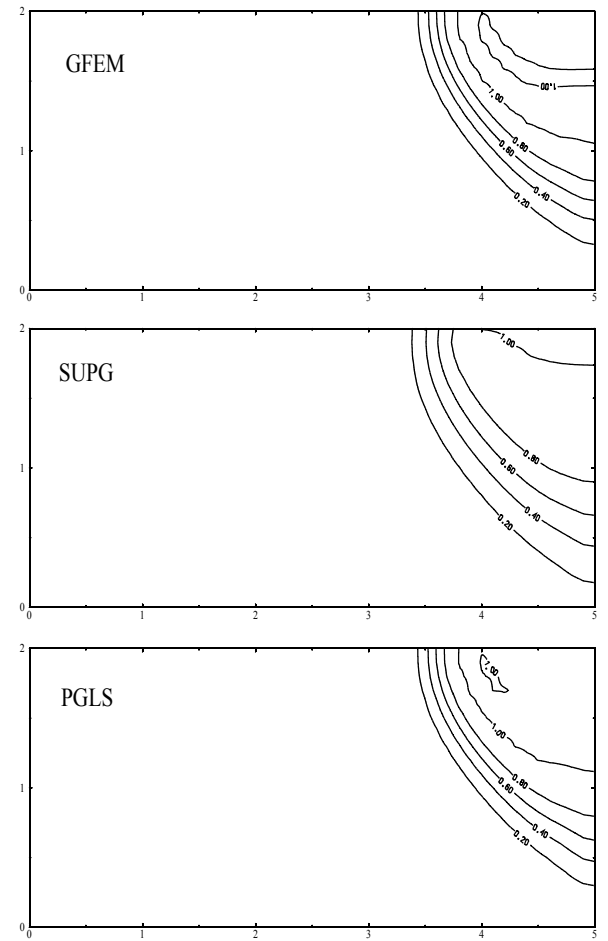
Figure 13.38 Concentration distribution after 10,000 days for the GFEM, SUPG and PGLS schemes of FEFLOW.

13.5 Conclusions

We benchmarked the PGLS technique which has been implemented into the FEFLOW package for transient 2D and 3D problems. The features and results we found give rise to the following conclusions:

(1) The PGLS represents an alternative numerical scheme to solve transient advection-dispersion transport problems. In contrast to standard techniques PGLS leads to symmetric matrix systems and possesses a built-in streamline-like upwind characteristics without any free parameter.

(2) The PGLS produces results which are comparable to the standard techniques if the processes are characterized by low to moderate advection.

(3) For advection-dominant problems (high Pg numbers) the PGLS cannot become wiggle-free.

(4) At high Pg numbers (coarse mesh) and small Co numbers (small time steps) the PGLS can run into difficulties. This seems to be a 'natural' property of the PGLS, because the damping matrix G in (13-63) is weighted by the time step size. If decreasing Δt the damping matrix G reduces with a quadratic descending rate, while the other matrix terms in (13-60) or (13-61) possesses only a linear or an independent descends of Δt. As the result, the PGLS scheme must become instable for sufficiently small Δt.

(5) The PGLS damping matrix G in (13-63) is strongly related to the pore velocity $q/[\varepsilon s(\psi)]$. For dry unsaturated problems with $s(\psi) \to 0$ the matrix G becomes singular (*cf.* remark 2).

(6) The PGLS is not applicable to the divergence form of the governing transport equation. On the other hand, the maintenance of PGLS's symmetric matrix property forces to drop any other techniques than simple Picard-iteration-based procedures for nonlinear problems. Here, Newton iteration would not be applicable for the PGLS (unless the symmetry is given up in favor of a better convergence of the Newton scheme).

From the practical point of view, the PGLS provides an additional alternative in modeling advection-dispersion equations. The advantage coming with the PGLS is in form of a symmetric matrix system.

References

1. Burnett, R. D. & Frind, E O., Simulation of contaminant transport in three dimensions, 1. The alternating direction Galerkin technique. *Water Resour. Res.* **23** (1987)4, 683-694.
2. Burnett, R. D. & Frind, E O., Simulation of contaminant transport in three dimensions, 2. Dimensionality effects. *Water Resour. Res.* **23** (1987)4, 695-705.
3. Codina, R., On stabilized finite element methods for linear systems of convection-diffusion-reaction equations. *Comput. Methods Appl. Mech. Engrg.* **188** (2000), 61-82.
4. Diersch, H.-J. G. Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.
5. Diersch, H.-J, G. A shock-capturing finite-element technique for unsaturated-saturated flow and transport problems. in: V. N. Burganos *et al.* (Eds.), *Computational Methods in Water Resources XII*, vol. 1, Computational Mechanics Publications, Southampton, 1998, 207-214.

6. Diersch H.-J. G. & Perrochet, P. On the primary variable switching technique for simulating unsaturated-saturated flows. *Adv. Water Resour.* **23** (1999), 25-55.

7. Donea, J., Guiliani, S., Laval, H. & Quartapelle, L., Time-accurate solution of advection-diffusion problems by finite elements. *Comput. Methods in Appl. Mech. and Engrg.* **45** (1984), 123-145.

8. Hoopes,J. A. & Harlemann, D. R. F., Wastewater recharge and dispersion in porous media. *ASCE J. Hydr. Division* **93** (1967) HY5, 51-71.

9. Hughes, T. J. R. & Brooks, A. N., A multi-dimensional upwind crosswind diffusion. in *Finite Element Methods for Convection Dominated Flows*, AMD **Vol 34**, ASME, Hughes T. J. R. ed., New York, 1979

10. Hughes, T. J. R., Mallet, M. & Mizukami, A., A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Comput. Meths. Appl. Mech. Engrg.* **54** (1986), 341-355.

11. Huyakorn, P. S., Andersen, P. F., Güven, P. F. & Molz, F. J., A curvi-linear finite element model for simulating two-well tracer tests and transport in stratified aquifers. *Water Resour. Res.* **22** (1986)5, 663-678.

12. Kelly, D. W., Nakazawa, S., Zienkiewicz, O. C. & Heinrich, J. C., A note on upwinding and anisotropic balancing in finite element convective-diffusion problems. *Int. Num. Meth. Engng.* **15** (1980), 1705-1711.

13. König, C., Operator split for three dimensional mass transport equation. *Proceedings of the Tenth Intern. Conf. on Computational Methods in Water Resources, Peters A.., Wittum G., Herrling B., Meissner U., Brebbia C.A., Gray W.G, Pinder, G.F, Vol. 1, Kluwer Academic Publ., Dordrecht 1994, 309-316.*

14. Leismann, H. M. & Frind, E. O., A symmetric-matrix time integration scheme for the efficient solution of advection-dispersion problems. *Water Resour. Res.* **25** (1989)6, 1133-1139.

15. Nguyen, H. & Reynen, J., A Space-time least-squares finite advection-diffusion equations. *Comput. Methods in Appl. Mech. and Engrg.* **42** (1984), 331-342.

16. Nguyen, H., A two-dimensional Petrov-Galerkin finite element model for the analysis of contaminant transport in saturated-unsaturated porous media. *Technical Note,* **N I.97.94**, JRC Ispra, June 1997.

17. Segol, G., *Classic groundwater simulations - Proving and improving numerical models*. PTR Prentice Hall, Englewood Cliffs, NJ, 1994.

18. Wendland, E. C., Numerische Simulation von Strömung und hochadvektivem Stofftransport in geklüftetem, porösem Medium. *Dissertation, Mitteilung Nr. 96-6*, Ruhr-Universität Bochum, 1996.

Appendix A

Nomenclature

Latin symbols

C, C_o	ML^{-3}	concentration and reference concentration of a miscible chemical species, respectively;
c^f, c^s	$L^2T^{-2}\Theta^{-1}$	specific heat capacity of fluid and solid, respectively;
$\boldsymbol{D}$	L^2T^{-1}	tensor of mechanical dispersion;
D_d	L^2T^{-1}	molecular diffusion in the porous medium;
e	l	gravitational unit vector;
g	LT^{-2}	gravitational acceleration;
h	L	hydraulic (piezometric) head;
$\boldsymbol{I}$	l	unit tensor;
$\boldsymbol{K}$	LT^{-1}	tensor of hydraulic conductivity for the saturated medium (anisotropy);
$\boldsymbol{k}$	L^2	tensor of permeability for the saturated medium (anisotropy);
K_r	l	relative hydraulic conductivity ($0 < K_r \le 1$, $K_r = 1$ if saturated at $s = 1$);
N_i		finite element shape function at node i;
p	$ML^{-1}T^{-2}$	fluid pressure;
Q_h	T^{-1}	fluid flow sink/source;
Q_C	$ML^{-3}T^{-1}$	bulk mass sink/source;

Q_T	$ML^{-1}T^{-3}$	bulk thermal sink/source;
Q_T^f, Q_T^s	$ML^{-1}T^{-3}$	fluid and solid thermal sink/source, respectively;
$\mathbf{q}$	LT^{-1}	Darcy flux vector;
q_n		normal flux on a boundary (positive outward);
R, R_d	1	retardation and derivative retardation, respectively;
S_o	L^{-1}	storage coefficient;
s	1	saturation of the fluid phase ($0 < s \leq 1$, $s = 1$ if medium is saturated);
s_r	1	residual saturation;
s_s	1	maximum saturation;
T, T_o	Θ	temperature and reference temperature, respectively;
v	LT^{-1}	$q/(\varepsilon s)$ pore velocity;
$\mathbf{v}$	LT^{-1}	$\mathbf{q}/(\varepsilon s)$ pore velocity vector;
z	L	elevation above a reference datum;

Greek symbols

α	$L^3 M^{-1}$	solutal expansion coefficient;
β	Θ^{-1}	thermal expansion coefficient;
β_L, β_T	L	longitudinal and transverse dispersivity, respectively;
Γ		boundary;
γ	L^{-1}	fluid compressibility;
Δl	L	characteristic element length;
Δt_n	T	time increment at level n;
ε	1	porosity ($0 < \varepsilon \leq 1$);
ϑ	T^{-1}	chemical decay rate;
Δ	$MLT^{-3}\Theta^{-1}$	tensor of thermal hydrodynamic dispersion of fluid phase;
λ^f, λ^s	$MLT^{-3}\Theta^{-1}$	thermal conductivity for fluid and

		solid, respectively;
μ, μ_o	$ML^{-1}T^{-2}$	dynamic viscosity and reference dynamic viscosity of fluid, respectively;
ρ, ρ_o	ML^{-3}	fluid density and reference fluid density, respectively;
ρ^s	ML^{-3}	solid density;
ϕ	L^{-1}	skeleton compressibility;
χ		adsorption function to describe Henry, Freundlich and Langmuir isotherms;
ψ	L	pressure head ($\psi > 0$ saturated medium, $\psi \leq 0$ unsaturated medium);
Ω		domain;

Subscripts

i, j, k	nodal indices;
o	reference value;

Superscripts

e	element;
f	fluid (water) phase;
n	time level;
s	solid phase;

Abbreviations

AB/TR	Adams-Bashforth/trapezoid rule predictor-corrector technique;
Co	Courant number;
FE/BE	forward Euler/backward Euler predictor-corrector technique;

13. The Petrov-Galerkin least square method (PGLS)

GFEM	Galerkin-based finite element method;
Pg	grid Peclet number;
PGLS	Petrov-Galerkin least-square;
SUPG	streamline-upwind Petrov-Galerkin

Extended formulations of constraints for Cauchy-type (3rd kind) boundary conditions in FEFLOW

14

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

14.1 Basic Formulation of 3rd Kind Boundary Conditions

FEFLOW provides alternative formulations of constraints for Cauchy-type (3rd kind) flow boundary conditions (BC's). Cauchy-type BC's are commonly used to describe river and other surface water boundaries in groundwater modeling. Their mathematical formulation is given by the following expression[1] written for groundwater flow (*cf.* Fig. 14.1)

$$q_n(\boldsymbol{x}, t) = -\Phi(h^R - h) \qquad (14\text{-}1a)$$

with

$$\Phi = \begin{cases} \Phi^{\text{in}} & \text{for} \quad h^R > h \\ \Phi^{\text{out}} & \text{for} \quad h^R \le h \end{cases} \qquad (14\text{-}1b)$$

where the parameters and variables can be temporarily and spatially dependent:
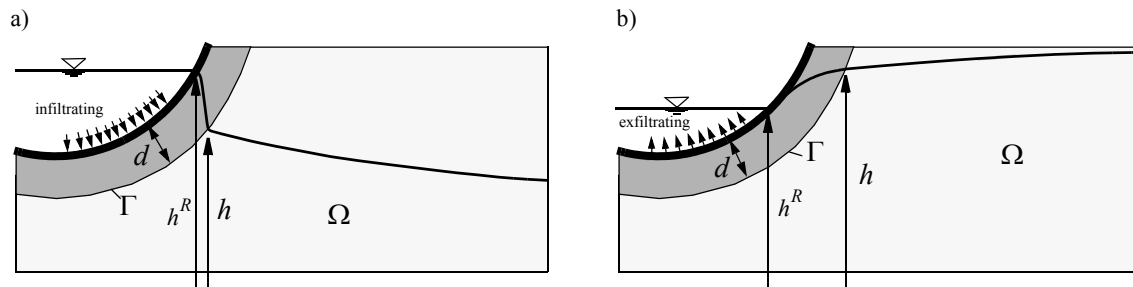
Figure 14.1 Transfer through a clogged river bed for a) infiltrating and b) exfiltrating conditions.

$$
\left.\begin{aligned}
\Phi^{in} &= \Phi^{in}(x, t) \\
\Phi^{out} &= \Phi^{out}(x, t) \\
h &= h(x, t) \\
h^R &= h^R(x, t)
\end{aligned}\right\} \qquad (14\text{-}1c)
$$

in which

d	=	thickness of the clogged river bed ('colmation' zone), $[L]$;
h	=	hydraulic head, $[L]$;
h^R	=	reference hydraulic head (e.g., water level of the river), $[L]$;
q_n	=	normal Darcy flux going through the boundary (positive outward), $[LT^{-1}]$;
t	=	time, $[T]$;
x	=	$\{x, y, z\}$ space coordinates, $[L]$;
Γ	=	boundary of groundwater domain Ω ;
Φ	=	transfer coefficient (colmation, leakage), $[T^{-1}]$;
Φ^{in}	=	directional in-transfer coefficient, $[T^{-1}]$;
Φ^{out}	=	directional out-transfer coefficient, $[T^{-1}]$;
Ω	=	groundwater modeling domain;

14.2 Optional Constraints for 3rd Kind Boundary Conditions

14.2.1 The standard form

The formulation of constraints is based on the formalism of *complementary conditions* for a type of BC[1]. Cauchy BC's are related to a potential condition which has to be prescribed on the boundary Γ in form of the reference hydraulic head h^R. Naturally, such type of BC is to be constrained by maximum and minimum fluxes, Q^{max}, Q^{min}. The following standard form of constraints for 3rd kind BC's is used:

$h^R(t)$ is valid if the flux is in the bounds: $\qquad (14\text{-}2)$

$$
\begin{cases}
\text{Max: } Q^R < Q^{max}(t) & \text{else} \quad Q^R = Q^{max}(t) \\
\text{Min: } Q^R > Q^{min}(t) & \text{else} \quad Q^R = Q^{min}(t)
\end{cases}
$$

where $Q^R = -\int q_n dV \; [L^3 T^{-1}]$ represents the summed-up, called *lumped* balance fluxes at nodal points to which the Cauchy-type boundary values are related. A typical example of a flux-limiting infiltration from a river bed is shown in the sketch of Fig. 14.2. If the groundwater table decreases below the location of the river bed a specific situation in form of a 'flow separation' occurs. Physically, the zone between the river bed and the water table becomes unsaturated and the linear relationship (14-1a) for the infiltrating water as a function of the difference Δh between the reference (river) water head h^R and the groundwater head h cannot be maintained anymore. It requires the prescription of the maximum bound Q^{max}.

Figure 14.2 Flux-limiting infiltration from a river bed formulated by a maximum flux constraint $Q^{max}(t)$.
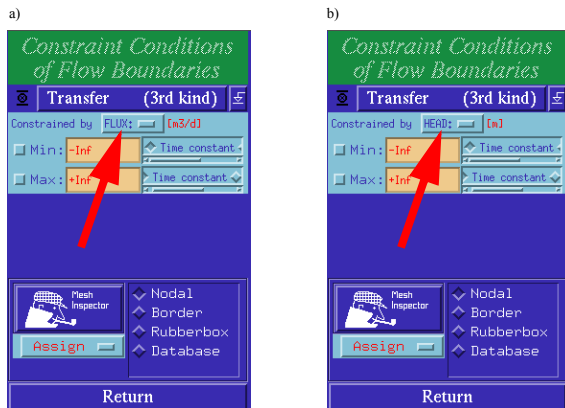
Figure 14.3 Input menu for prescribing a) flux-constrained and b) head-constrained transfer conditions for Cauchy-type BC's.

FEFLOW provides the standard form of constraints for Cauchy-type BC's, where the bounds Q^{max} and/or Q^{min} are directly input as shown in Fig. 14.3a. This formulation is termed as *flux-constrained transfer* BC.

Such a limitation by fluxes represents a quite general formulation. The min-max fluxes are user-specified input parameters, which can be quantified by user-own rules and data. However, the disadvantage is here that the determination of the constraint fluxes requires geometric information of the boundaries (e.g., transfer areas) at a given mesh. The new formulation of constraints to be described next overcomes this drawback and is more useful in practice.

14.2.2 The new form

Instead of prescribing the constraint fluxes directly the new form allows the input of maximum and minimum head values, h^{max}, h^{min}, which are used to derive the constrained min-max fluxes for Cauchy-type BC's. This form is termed *head-constrained transfer* BC and can be optionally selected in the input menu for transfer constraint conditions of flow boundaries (Fig. 14.3b). In contrast to Eq. (14-2) the mathematical formulation of the new head-constrained transfer BC is as follows:

$h^R(t)$ is valid if the head is in the bounds: (14-3)

$$
\begin{cases}
\text{Max: } h < h^{max}(t) & \text{else} \quad q_n = q_n^{max} = -\Phi(h^R - h^{max}) \\
\text{Min: } h > h^{min}(t) & \text{else} \quad q_n = q_n^{min} = -\Phi(h^R - h^{min})
\end{cases}
$$

The advantage of this constraint formulation is that the limiting (constraint) fluxes q_n^{max}, q_n^{min} are rates (no more discharges!), which are computed from the input head constraints h^{max}, h^{min}. The physical meaning of the head constraints is explained for the minimum head h^{min} limit as illustrated in Fig. 14.4.
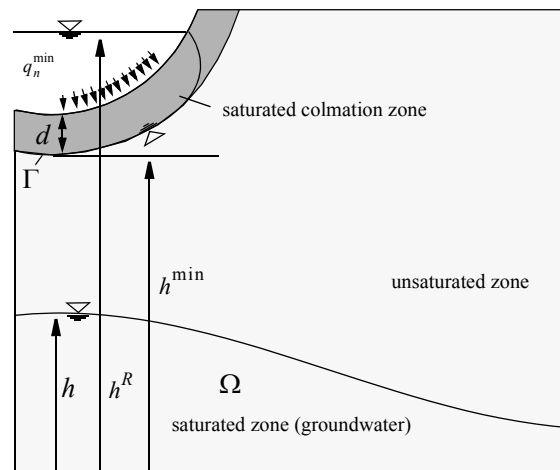
Figure 14.4 New head-constrained transfer BC for a flux-limiting infiltration from a river bed**.**

The river bed is clogged by a colmation zone with a thickness d. A perched water situation occurs, where the groundwater table has no more a direct hydraulic contact with the surface water. An unsaturated zone forms below the river bed. Typically, in the colmation zone the subsurface water remains saturated (Fig. 14.4). Assuming the validity of Darcy's law for the colmation zone, the normal flux from the river entering the unsaturated zone can be assessed as

$$q_n = -K^{col}\frac{\Delta h}{\Delta z} = -\frac{K^{col}}{d}(h^R - h^{min}) \qquad (14\text{-}4)$$

where K^{col} is the saturated hydraulic conductivity of the colmation zone. It can be easily seen that Eq. (14-4) is equivalent to the constraint flux q_n^{min} in Eq. (14-3) derived from the minimum head limit h^{min}, where the transfer coefficient is simply given by

$$\Phi = \frac{K^{col}}{d} \qquad (14\text{-}5)$$

REMARK 1: The directional in-transfer and out-transfer coefficients Φ^{in}, Φ^{out} are assigned according to the minimum and maximum head constraints, *viz.*,

$$\Phi = \begin{cases} \Phi^{in} & \text{for} \quad h^{min} \quad \text{since} \quad h^R > h^{min} \\ \Phi^{out} & \text{for} \quad h^{max} \quad \text{since} \quad h^R \le h^{max} \end{cases} \qquad (14\text{-}6)$$

REMARK 2: The new head-constrained transfer BC is very efficient. It need not a switching of BC's if constraints are set and reset during a simulation run.

REMARK 3: Time-dependent head-constraints are appropriate to prescribe intermediate flux conditions along a boundary (e.g., at certain times no flux conditions should occur as applied to temporarily moving BC's). Since $h^R = h^R(t)$ a temporal no flux condition is automatically satisfied if the reference head h^R becomes identical to the constrained head h^{min} (or h^{max}) in time, that means written for the minimum constraint

$$q_n = q_n^{min} \equiv 0 \quad \text{for} \qquad\qquad (14\text{-}7)$$
$$h^R(t) = h^{min}(t) \quad \text{and} \quad h(t) < h^{min}(t)$$

To force a temporal no flux conditions independent of the groundwater head h, the maximum head constraint has to be set additionally to the reference head. It requires

$$q_n = q_n^{min} \equiv 0 \quad \text{for} \qquad\qquad (14\text{-}8)$$
$$h^R(t) = h^{min}(t) = h^{max}(t) \quad \text{and arbitrary} \quad h(t)$$

References

1. Diersch, H.-J.G. Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY, Berlin, 2002.

14. Extended formulations of constraints for Cauchy-type (3rd kind) boundary conditions in FEFLOW

15

Nonlinear dispersion in density-dependent mass transport

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

15.1 Introduction

In modeling density-dependent flow and mass transport problems an increasing interest has been cases where high-concentration differences in the system occur, e.g., applications to hazardous waste disposal in salt formations or brine transport in deep aquifers. Traditionally, density-dependent mass transport is modeled on the basis of the classic Darcy law and the linear Fickian dispersion equation. But, in one-dimensional laboratory experiments [2,4] with high-concentration gradients it was found that the dispersivity does not seem to be a property of the porous medium alone. It was observed that the mixing process of salt-water is dependent on the concentration gradient and the dispersivity had to be changed from case to case to get a sufficient fit to the measurements. Using same porous media the dispersivity had to be decreased as the difference in concentration of the resident and displacing fluids increased. In past, various attempts were made to explain this phenomenon. A formal dependence of dispersivities on the salt concentration has shown an inappropriate and a theoretically contrary approach because the dispersivities are a geometric property of the porous medium and should not be dependent on the physicochemical property of the fluid flowing through the voids.

Hassanizadeh & Leijnse[2] and Hassanizadeh[3] have proposed extensions of the dispersion theory in form of a non-Fickian law. In using such a nonlinear dispersion theory the laboratory experiments could be explained and fit reasonably. New experiments have confirmed these theoretical findings[4]. Furthermore, from the theoretical point of view the non-Fickian dispersion is consistent with the classic approach and theoretically well founded.

The nonlinear (non-Fickian) dispersion law has been implemented in FEFLOW[1] both for 2D and 3D mass transport processes. It represents an extension to the classic linear Bear-Scheidegger dispersion law and can be optionally selected. In the following the theoretical basis of the nonlinear dispersion theory will be described in some detail. Then the implementation in FEFLOW and the numerical solution of the nonlinear mass fluxes will be discussed. Finally, an example is presented which benchmarks the implemented rules.

15.2 Basic Equations

The starting point forms the mass conservation equation written as

$$\frac{\partial(\varepsilon C)}{\partial t} + \nabla \cdot (C\boldsymbol{q} + \boldsymbol{J}) = r \tag{15-1}$$

where the symbols are summarized below in the section 'Notation'. The dispersive mass flux vector $\boldsymbol{J}$ is commonly expressed by the *linear Fickian type equation*, *viz.*,

$$\boldsymbol{J} = -\boldsymbol{D} \cdot \nabla C \tag{15-2}$$

The hydrodynamic dispersion tensor $\boldsymbol{D}$ is assumed to be independent of the concentration C and its gradient. It is, however, considered to be a function of the flow velocity $\boldsymbol{q}$ and is commonly described by the *Bear-Scheidegger dispersion* relationship for a porous medium according to

$$\boldsymbol{D} = (\varepsilon D_d + \beta_T \|\boldsymbol{q}\|)\boldsymbol{I} + (\beta_L - \beta_T)\frac{\boldsymbol{q} \otimes \boldsymbol{q}}{\|\boldsymbol{q}\|} \tag{15-3}$$

For the linear Fickian law (15-2) the dispersive mass flux of a solute is proportional to the solute concentration gradient. But, if large concentration gradients exist, nonlinear effects become important and $\boldsymbol{J}$ has to be replaced by an extended *nonlinear (non-Fickian) dispersion law*[2]

$$\boldsymbol{J}(\beta\|\boldsymbol{J}\| + 1) = -\boldsymbol{D} \cdot \nabla C \tag{15-4}$$

where β is a new high-concentration (HC) dispersion coefficient and $\boldsymbol{D}$ is still the known Bear-Scheidegger dispersion tensor given by Eq. (15-3) with longitudinal and transverse dispersivities considered to be (constant) properties of the porous medium and independent of the fluid properties and transport processes.

High concentration-gradient experiments[2,4] have shown that the nonlinear dispersion law (15-4) gives very good fits to measured breakthrough curves. It is found that the HC dispersion coefficient β varies inversely with the flow velocity $\boldsymbol{q}$. Schotting *et al.*[4] have summarized their fitted experiments in the following approximate expression for $\beta = \beta(\boldsymbol{q})$ as

$$\beta(\boldsymbol{q}) = \frac{0.0125}{\|\boldsymbol{q}\|^{1.76}}[\text{s m}^2 / \text{kg}] \qquad \text{for} \tag{15-5}$$

$$9 \cdot 10^{-5} < \|\boldsymbol{q}\| < 3 \cdot 10^{-3}[\text{m}/\text{s}]$$

15.3 Implementation in FEFLOW

15.3.1 Selection of dispersion laws

In FEFLOW the modeler can use either the standard linear Fickian dispersion law (15-2) or the nonlinear non-Fickian dispersion law (15-4). Both options can be set in the Problem Editor for the mass material data. Figure 15.1 exhibits this part of the editor, where the both dispersion laws can be chosen and the additional HC-β-parameter is input in the case of the nonlinear dispersion.
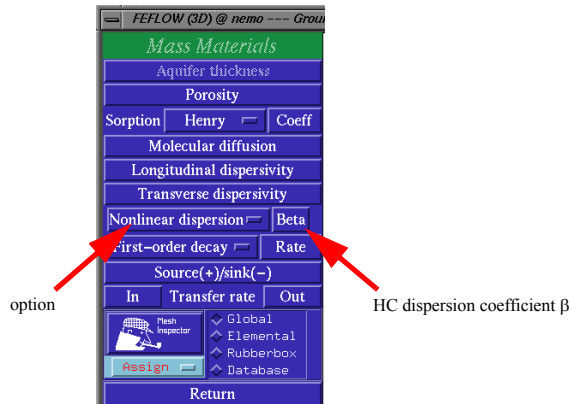
Figure 15.1 FEFLOW's mass material data editor menu to choose the mass dispersion law option and to input the HC dispersion coefficient in the case of nonlinear dispersion.

$$
\left.
\begin{array}{lll}
0.\ \text{initial} & J_0 = 0 \\
1.\ \text{step} & J_1 = -\dfrac{D}{\beta\|J_0\| + 1} \cdot \nabla C_1 \\
2.\ \text{step} & J_2 = -\dfrac{D}{\beta\|J_1\| + 1} \cdot \nabla C_2 \\
\qquad\cdots \\
\tau.\ \text{step} & J_\tau = -\dfrac{D}{\beta\|J_{\tau-1}\| + 1} \cdot \nabla C_\tau
\end{array}
\right\}
\qquad (15\text{-}6)
$$

where τ represents an iteration counter. The iteration (15-6) is performed at each time step in dependence on the selected time stepping strategy: (1) For fixed (pre-defined) time steps it is iterated at each time level. The procedure is terminated if the convergence criterion is satisfied. (2) For the adaptive predictor-corrector time marching the nonlinear solution is fully controlled by the time step itself, where the nonlinear dispersion is linearized in time according to

$$
J_{n+1} = -\frac{D}{\beta\|J_n\| + 1} \cdot \nabla C_{n+1}
\qquad (15\text{-}7)
$$

where n corresponds to the time level.

15.3.2 Numerical solution for nonlinear dispersion

The numerical solution of the governing balance equation (15-1) with the nonlinear dispersion law (15-4) requires a specific iterative strategy. A recursive scheme is preferred which is performed by the following iteration procedure:

15.4 Example

Schotting *et al.*[4] have derived analytical solutions in 1D, which will be used to benchmark FEFLOW with the nonlinear dispersion law. We consider the displacement of a high concentration through a column with constant properties. The data are summarized in Tab. 15.1.

Table 15.1 Data of the displacement experiment

Quantity	Symbol	Magnitude	Unit
Length of column ($-0.5\,\mathrm{m} \leq z \leq 4.0\,\mathrm{m}$)	L	4.5	m
Flow rate	q_o	$3.209 \cdot 10^{-5}$ 2.772576	$\mathrm{m\,s^{-1}}$ $\mathrm{m\,d^{-1}}$
Porosity	ε	0.2	1
Boundary concentration (brine input)	C_s	$2.85714 \cdot 10^5$	$\mathrm{mg\,l^{-1}}$
Molecular diffusion	D_d	0.0	$\mathrm{m^2 s^{-1}}$
Longitudinal dispersivity	β_L	1.0	m
HC-dispersion coefficient	β	10^4 $1.1574 \cdot 10^{-4}$	$\mathrm{m^2 s\,kg^{-1}}$ $\mathrm{m^2 d\,g^{-1}}$

The column is initially filled with freshwater ρ_o, $C = 0$. At $t = 0$ brine $\rho_s = \rho_s(C_s)$ starts entering the column with uniform specific discharge q_o. This implies the following initial condition

$$C(z, 0) = \begin{cases} C_s & \text{for} \quad z < 0 \\ 0 & \text{for} \quad z \geq 0 \end{cases} \tag{15-8}$$

In FEFLOW's numerical simulation the outflowing boundary is imposed with a natural gradient boundary condition $\nabla C|_{z = 4\mathrm{m}} \approx 0$. The column is discretized by 900 linear quadrilateral elements resulting an spatial increment of $\Delta z = 0.005\,\mathrm{m}$. For the temporal approximation the default forward Adams-Bashforth/backward-trapezoid rule predictor-corrector scheme with

adaptive time stepping (error tolerance 10^{-4}) is used. It requires 144 time steps to simulate the displacement process for a dimensionless time Υ defined as

$$\Upsilon = t\frac{q_o^2}{\varepsilon \|D\|} \rightarrow \frac{tq_o}{\varepsilon \beta_L} \tag{15-9}$$

up to $\Upsilon = 1.0$. The numerical results are in a very good agreement with the analytical results obtained by Schotting et al.[4] in form of a semi-explicit solution as shown in Fig. 15.2.
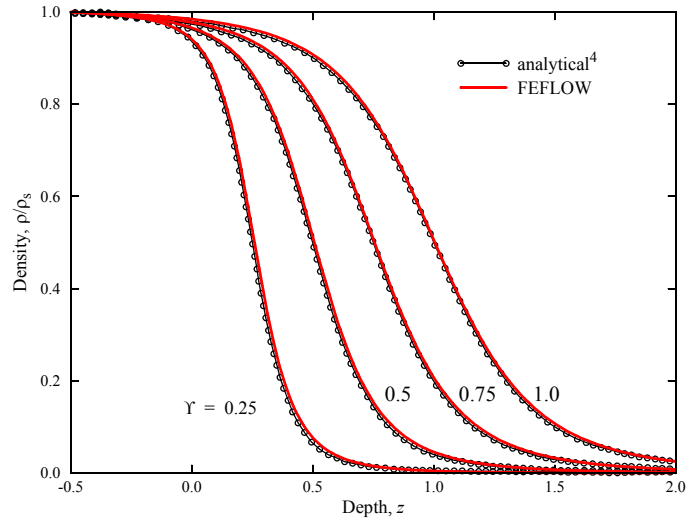
Figure 15.2 Numerical density profiles simulated by FEFLOW at selected dimensionless times Υ in comparison with the semi-explicit analytical solutions given by Schotting et al.[4] for a brine displacement in a column at nonlinear dispersion.

Notation

C	=	salinity, saltwater concentration, (ML^{-3}) ;
C_s	=	brine input concentration, (ML^{-3}) ;
D	=	tensor of hydrodynamic dispersion, $(L^2 T^{-1})$;
D_d	=	effective molecular diffusion, $(L^2 T^{-1})$;
I	=	unit (identity) tensor, (1) ;
J	=	dispersive mass flux vector, $(ML^{-2}T^{-1})$;
n	=	time level;
q	=	Darcy velocity vector, (LT^{-1}) ;

q_o	=	uniform specific discharge, (LT^{-1}) ;
r_c	=	chemical reaction rate, $(ML^{-3}T^{-1})$;
t	=	time, (T) ;
β	=	high-concentration (HC) dispersion coefficient, $(M^{-1}L^2 T)$;
β_L, β_T	=	longitudinal and transverse dispersivity, respectively, (L) ;
ε	=	porosity, (1) ;
ρ_o	=	density of freshwater, (ML^{-3}) ;
ρ_s	=	brine density, (ML^{-3}) ;
τ	=	iteration counter;
Υ	=	dimensionless time, (1) ;
∇	=	Nabla (vector) operator, (L^{-1}) ;

References

1. Diersch, H.-J. G., Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.

2. Hassanizadeh, S.M. & Leijnse, A., A non-linear theory of high-concentration-gradient dispersion in porous media. *Adv. Water Resources* **18** (1995) 4, 203-215.

3. Hassanizadeh, S.M., On the transient non-Fickian dispersion theory. *Transport in Porous Media* **23** (1996), 107-124.

4. Schotting, R.J., Moser, H. & Hassanizadeh, S.M., High-concentration-gradient dispersion in porous media: experiments, analysis and approximations. *Adv. Water Resources* **22** (1999) 7, 665-680.

Consistent velocity approximation in the finite-element simulation of density-dependent mass and heat transport processes

16

H.-J. G. Diersch

WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany

16.1 Introduction

To compute the derivatives (velocities) from a finite element approximation, the obvious and self-consistent approach is that of directly differentiating the finite element solution at points of interests. In groundwater flow simulation local Darcy velocities are calculated by differentiating the hydraulic head h (or pressure p) solution and using the appropriate constitutive relations. This direct computation results in lower order, discontinues derivatives with inferior accuracy at the boundary of the elements and at the interelement nodes where accurate values of fluxes are usually desired. To achieve accurate derivatives different techniques are practiced.

In the finite element method local and global smoothing (*projection*) techniques are commonly applied where the derivatives are computed at optimal sampling (Gauss) points. The local projection procedure consists of an extrapolation from the superconvergent points and a subsequent averaging at nodes to obtain accurate nodal velocity values. Global smoothing assume a continues interpolation of the derivatives

of the same form as that used for the basis functions. Such kind of projections are considered as *consistent finite element derivatives*. FEFLOW uses these techniques which are described in Diersch and Kolditz[6] in the context of coupled flow and transport simulation, which are summarized in the Appendix B. In practical finite element computations and numerous benchmark tests[3,4,6,14] it has been shown that continuous derivatives by using the velocity smoothing techniques give accurate solutions.

In density-dependent flow and transport processes a proper care should be taken in the derivation of the velocities. This has to do with the lower-order approximation attainable for the pressure (or head) gradients ∇p which can conflict the high-order spatial variability in the gravity (buoyancy) term ρg. The problem has been addressed by Voss[20], Voss and Souza[21], Herbert *et al.*[11] and Leijnse[17], who proposed modified schemes in evaluation the discontinuous derivatives termed as *consistent velocity approximation*. In Voss and Souza's approach the spatial variation in the gravity term is reduced to the same spatial functionality occurred for the pressure gradient, i.e., for linear finite elements the pressure gradient is constant (piecewise constant per

element) and accordingly the gravity term should be also piecewise constant. Leijnse extends this procedure and prefers averages of the buoyancy term only in the appropriate gravity direction. While Voss and Souza[21] and Leijnse[17] tried to overcome the problem of consistency by precision reduction, Herbert et al.[11] solved it by introducing a second-order approximation for the pressure (quadratic shape functions) and a linear-order approximation for the salinity (linear shape functions). Herbert et al.'s mixed approach is more natural in the finite element method (weighted Galerkin statements), but, it results in additional computational costs.

The smoothing techniques[6] (local and global projections) for the velocities which are derived in the elements without modifying the spatial variability in the gravity term (no precision reduction) are an efficient alternative for deriving consistent velocity fields. It results in a continuous representation of the nodal velocities. The derivatives are used on an element patch surrounding the nodes. In this way the technique is consistent with an integral evaluation of the flux terms and the patch-related nodal velocities represents averaged quantities in the weighted sense.

The question arises now whether the projection (smoothing) techniques for deriving patch-related nodal velocities are sufficiently consistent for density-dependent problems. Indeed, we can confirm it in all previous tests and applications. The benchmark tests in form of the Henry problem[4,14], the Elder problem[6,14], salt dome problem[6,14] and others[5,7,13] revealed good agreements with the other solutions available. More recently, however, an obvious counter-example of a high-contrast density problem gives rise to a critical review of the used velocity derivation procedures.

Oswald[18] performed a series of three-dimensional laboratory experiments termed as the *saltpool* problem which involves stable layering of saltwater below freshwater. A discharge of water causes a transient upconing of saltwater. In varying the density contrast in the upconing process the measured quantities for instance in form of the salinity breakthrough at the outlet have shown a significant influence due to gravity effects. Oswald et al.[19] tried to recompute the saltpool processes by using different codes (for more see[18]). While at a lower salinity (1% salt mass fraction) a good agreement was found, at high salinity (10% salt mass fraction) the computed saltwater concentration at the outlet became generally too large and often quite depart from the measurements. This discrepancies gave rise to numerous investigations and the development of improved numerical schemes.

Ackerer et al.[2] applied a new numerical code TVDV-3D which is based on mixed and discontinuous finite elements (for more details see Ackerer et al.[1]). Their results are satisfactory, however, the simulation overestimates the saltwater mixing concentration for the lower density case and underestimates the saltwater breakthrough at the outlet for the high density case. Most recently, Johannsen et al.[16] presented improved results and achieved a good agreement with the measurements for both cases of low and high densities. Some parameters, particularly the permeability, the porosity and the transverse dispersivity were adjusted within accepted bounds in order to fully match the results of the physical experiment. From numerical point of view, their results are simulated with the new code d^3f (for more, see Frolkovic[9]) in which improved techniques for computing consistent velocities under

strong density effects are incorporated. These techniques were substantially responsible for the success of the difficult simulations of the saltpool problem at high salinity contrast.

As a consequence from the saltpool simulations the techniques for approximating the consistent velocities were also revised for the FEFLOW code. In the following a new improved technique is described and tested which is implemented in the FEFLOW code.

16.2 Basic Equations

For simulating density-coupled flow and transport processes the following equations have to be solved (symbols are listed in Appendix A):

$$\left.\begin{array}{c} \dfrac{\partial(\varepsilon\rho)}{\partial t} + \nabla \cdot (\rho\boldsymbol{v}) = Q_\rho \\[2mm] \dfrac{\partial(\varepsilon C)}{\partial t} + \nabla \cdot (C\boldsymbol{v} + \boldsymbol{J}_C) = Q_C \\[2mm] \dfrac{\partial}{\partial t}[\varepsilon\rho E^f + (1-\varepsilon)\rho^s E^s] + \nabla \cdot (\rho E^f \boldsymbol{v} + \boldsymbol{J}_E) = Q_E \end{array}\right\} \quad (16\text{-}1)$$

where the fluid velocity (Darcy flux) $\boldsymbol{v}$ is explicitly given by

$$\left.\begin{array}{c} \boldsymbol{v} = -\dfrac{\boldsymbol{k}}{\mu} \cdot (\nabla p - \rho\boldsymbol{g}) \\[2mm] \boldsymbol{v} = -\boldsymbol{K} f_\mu \cdot (\nabla h + \tilde{\rho}\boldsymbol{e}) \end{array}\right\} \quad (16\text{-}2)$$

introducing the relative density $\tilde{\rho}$

$$\tilde{\rho} = \frac{\rho - \rho_o}{\rho_o} \quad (16\text{-}3)$$

which is written in the two equivalent formulations in accordance with the chosen primary variable in form of the pressure p or the hydraulic head h, where the following relationships hold[4]

$$\left.\begin{array}{l} h = \dfrac{p}{\rho_o g} + z \\[2mm] \boldsymbol{g} = -g\boldsymbol{e} \\[2mm] f_\mu = \dfrac{\mu_o}{\mu} \\[2mm] \boldsymbol{K} = \dfrac{\boldsymbol{k}\rho_o g}{\mu_o} \\[2mm] E^f = E^f(T), \quad E^s = E^s(T) \\[2mm] \mu = \mu(C, T) \\[2mm] \rho = \rho(C, T) = \rho_0\Big[1 + \dfrac{\alpha}{C_s - C_o}(C - C_o) - \beta(T - T_o)\Big] \\[2mm] \boldsymbol{J}_C = -(\varepsilon D_d \boldsymbol{I} + \boldsymbol{D}) \cdot \nabla C \\[2mm] \boldsymbol{J}_E = -[(\varepsilon\lambda^f + (1-\varepsilon)\lambda^s)\boldsymbol{I} + \rho c \boldsymbol{D}] \cdot \nabla T \\[2mm] \boldsymbol{D} = (\beta_L - \beta_T)\dfrac{\boldsymbol{v}\otimes\boldsymbol{v}}{\|\boldsymbol{v}\|} + \beta_T\|\boldsymbol{v}\|\boldsymbol{I} \end{array}\right\} \quad (16\text{-}4)$$

The velocity $\boldsymbol{v}$ in (16-2) is coupled with the conservation equations (16-1) through the dependence of fluid density $\rho = \rho(C, T)$ on the concentration C and temperature T and, additionally, the dependence of fluid viscosity $\mu = \mu(C, T)$ on the concentration C and temperature T.

$$\rho = \rho_o + (\rho_1 - \rho_o)z \qquad z \in [0, 1] \qquad (16\text{-}5)$$

16.3 The Hydrostatic Condition

16.3.1 Equilibrium requirement: The requirement of consistency

Consider the following hydrostatic situation[12] for a finite element as shown in Fig. 16.1. For simplicity, the interval Δz is [0,1].

We assume the density ρ is varying linearly in the z-direction of gravity, *viz*.,

For the simple vertical problem the Darcy velocity is

$$\left.\begin{array}{l} v_x = -\dfrac{k_{xx}}{\mu}\partial_x p \\[2mm] v_y = -\dfrac{k_{yy}}{\mu}\partial_y p \\[2mm] v_z = -\dfrac{k_{zz}}{\mu}(\partial_z p + \rho g) \end{array}\right\} \qquad (16\text{-}6)$$
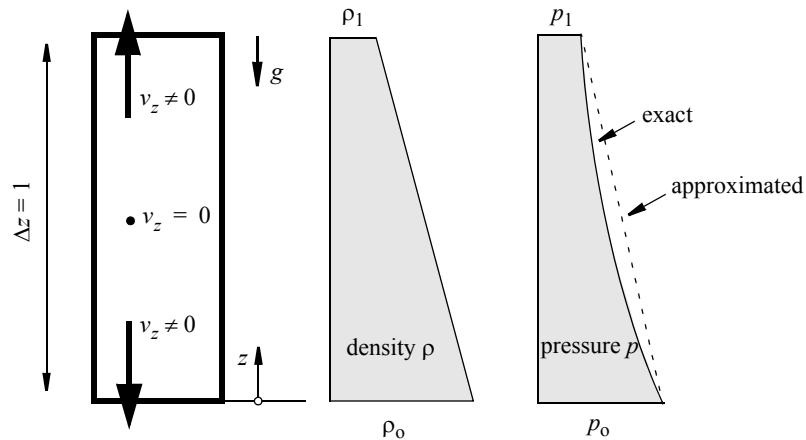
Figure 16.1 Hydrostatic conditions in a finite element of length $\Delta z = 1$ under a linear density gradient $\rho = \rho_o + (\rho_1 - \rho_o)z$; spurious vertical velocities v_z caused by an inexact pressure approximation.

Under a hydrostatic equilibrium the velocity vector v is (must be) zero everywhere. This is termed as the *requirement of consistency*:

$$v \equiv 0; \qquad \nabla p = \rho g \qquad \nabla h = -\tilde{\rho} e \qquad (16\text{-}7)$$

For the above example (Fig. 16.1) with a vertical density gradient we have to require

$$\partial_x p = \partial_y p = 0; \qquad \partial_z p = -\rho g \qquad \partial_z h = -\tilde{\rho} e_z \quad (16\text{-}8)$$

and the pressure p has to satisfy the following relationship

$$p = p(z) = -g\int_{z_o}^{z}\rho(\theta)d\theta \qquad (16\text{-}9)$$

which yields

$$p = p_0 - g\left(\rho_o z + \frac{\rho_1 - \rho_o}{2}z^2\right) \qquad (16\text{-}10)$$

Similar expression can be derived for the hydraulic head h:

$$h = h(z) = -e_z\int_{z_o}^{z}\tilde{\rho}(\theta)d\theta \qquad (16\text{-}11)$$

$$h = h_0 - e_z\left(\tilde{\rho}_o z + \frac{\tilde{\rho}_1 - \tilde{\rho}_o}{2}z^2\right) \qquad (16\text{-}12)$$

Equation (16-10) indicates that for a *linear* density ρ a *quadratic* shape of the pressure p is required to maintain a hydrostatic equilibrium for all z in the interval.

16.3.2 The artifact: Spurious nonconsistent velocities and common ways to overcome

Typically, in a discretization algorithm the concentration C and/or the temperature T is linearly approximated in a finite element. This leads to a corresponding linear relationship for the density ρ as considered above. But, the pressure p (or alternatively the head h) is also approximated by a linear function in an element. This is (in the example of Fig. 16.1)[12]

$$p \to \tilde{p} = p_o + (p_1 - p_o)z \qquad z \in [0, 1] \qquad (16\text{-}13)$$

Inserting (16-13) in the Darcy equation (16-2) or (16-6) and using the exact nodal values p_o and $p_1 = p_o - g(\rho_o + \rho_1)/2$ from (16-10) we get for the z-component of the approximated velocity

$$v_z = -\frac{k_{zz}}{\mu}g(\rho_o - \rho_1)\left(\frac{1}{2} - z\right) \qquad z \in [0, 1] \qquad (16\text{-}14)$$

It clearly indicates that the approximated velocity only vanishes at the middle point ($z = 1/2$) while at the other points *artificial nonzero quantities* occurs which take maximum values with opposite signs at the left and right point (*cf.* Fig. 16.1). Such *spurious nonconsistent velocities* can waste the computational results in form of an overestimation of the mixing processes at strong density coupling. In the advective terms of the governing transport equations (16-1) it will often not have a large effect, since the integration over elements and the assembly of adjacent elements averages out the nonconsistent velocities. However, if such spurious velocities are used to evaluate the dispersion tensor at element level an artificial increase of hydrodynamic dispersion (mixing) can result[17].

The most important way to overcome the problem is in reducing the spatial variability in the gravity term. Commonly, the gravity term is averaged in the appropriate direction so as proposed by Voss[20], Voss and Souza[21] and Leijnse[17]. In the above example we have to use now $\rho = (\rho_o + \rho_1)/2$ and find with the exact nodal

values p_o and $p_1 = p_o - g(\rho_o + \rho_1)/2$:

$$v_z = -\frac{k_{zz}}{\mu}\left(\underbrace{p_o - g\frac{(\rho_o + \rho_1)}{2} - p_o}_{\partial_z p} + \underbrace{g\frac{(\rho_o + \rho_1)}{2}}_{\rho g} \right) = 0 \qquad (16\text{-}15)$$

which satisfies the equilibrium at all points.

On other possibility is in averaging the nonconsistent velocities at nodal points by local or global projection (smoothing) techniques as mentioned above and thoroughly described in Appendix B. It smooths out the spurious velocities. Let us consider the following examples as shown in Fig. 16.2, where a node k is considered which is shared by two elements.
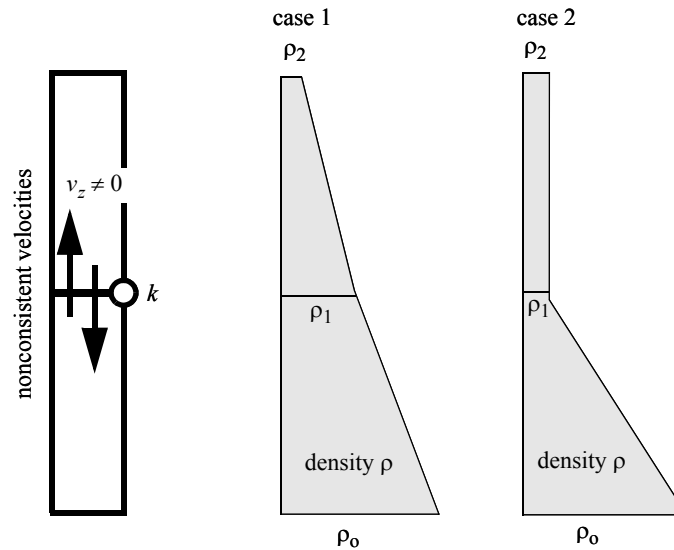
Figure 16.2 Continuous nodal velocity by averaging (smoothing) nonconsistent velocities for two cases of vertical density profiles.

The smoothing procedure for the nonconsistent velocity (16-14) leads to a velocity at the node k as

$$v_z = \frac{k_{zz}}{\mu}g\frac{1}{2}\left[\frac{(\rho_o + \rho_2)}{2} - \rho_1\right] \qquad (16\text{-}16)$$

If we can assume that the density ρ_1 at the node k is an average of the upper and lower density values, i.e., $\rho_1 = (\rho_o + \rho_2)/2$, then the nodal velocity (16-16) becomes consistent $v_z = 0$. Obviously, this is true (or approximately true) for typical density profiles as shown as case 1 in Fig. 16.2. However, if the density

profile is strongly variable over a short distance (e.g., a saltwater-freshwater interface with a high density contrast) the nonconsistent velocities do not average out. This can be seen for the case 2 in Fig. 16.2 at the node k where an upgoing spurious velocity remains in order of

$$v_z = \frac{k_{zz}}{\mu} g \frac{1}{2}\left[\frac{(\rho_o - \rho_1)}{2}\right] \qquad (16\text{-}17)$$

and the consistency is not satisfied at the node under such conditions.

We can summarize and conclude the following:
(1) Consistency is the requirement to a zero velocity under hydrostatic conditions for an arbitrary stable density gradient. A consistent velocity approximation satisfies the relationship (16-7) at the local evaluation points.

(2) Averaging of the gravity term for each element yields a consistent velocity approximation, however, the accuracy in the spatial variability is reduced.

(3) Smoothing of nonconsistent velocities derived at the Gaussian evaluation points averages out spurious velocities in the most cases. However, if the density gradients become very large spurious velocities at local points can remain. Accordingly, smoothing is a procedure to derive continuous nodal velocities which are often, but not always *consistent* in the sense of the statement (16-7).

(4) There is a desire to a more general, accurate and robust procedure for a consistent velocity approximation applied to density-dependent mass and heat transport problems.

16.4 New Formulation of Consistent Velocity

16.4.1 The improved Frolkovic and Knabner algorithm

Frolkovic[8,9] and Knabner and Frolkovic[12] proposed an new algorithm for approximating consistent velocities in two- and three-dimensional finite elements.

16.4.1.1 Transformations in local coordinates

The algorithm is described for affine and isoparametric families of elements, where the computations are realized on generalized (local) coordinates (ξ, η, ζ). The mapping from the local coordinates (ξ, η, ζ) to the global ones (x, y, z) is given by

$$\left.\begin{aligned} x &= x(\xi, \eta, \zeta) = \sum_m x_m N_m(\xi, \eta, \zeta) \\ y &= y(\xi, \eta, \zeta) = \sum_m y_m N_m(\xi, \eta, \zeta) \\ z &= z(\xi, \eta, \zeta) = \sum_m z_m N_m(\xi, \eta, \zeta) \end{aligned}\right\} \qquad (16\text{-}18)$$

where x_m, y_m, z_m, $m = 1, 2, \ldots$ are the coordinates of the vertices (nodes m) of the element and N_m are the finite element shape functions. The mapping requires that the transformation *Jacobian* $\boldsymbol{J}$ is nonsingular, where $\boldsymbol{J}$ is given by

16. Consistent velocity approximation in the finite-element simulation of density-dependent mass and heat transport processes

$$J = \begin{Bmatrix} \partial_\xi \\ \partial_\eta \\ \partial_\zeta \end{Bmatrix} \{x, y, z\} = \begin{bmatrix} \partial_\xi x & \partial_\xi y & \partial_\xi z \\ \partial_\eta x & \partial_\eta y & \partial_\eta z \\ \partial_\zeta x & \partial_\zeta y & \partial_\zeta z \end{bmatrix} \qquad (16\text{-}19)$$

$$= \sum_m \begin{bmatrix} (x_m \partial_\xi N_m) & (y_m \partial_\xi N_m) & (z_m \partial_\xi N_m) \\ (x_m \partial_\eta N_m) & (y_m \partial_\eta N_m) & (z_m \partial_\eta N_m) \\ (x_m \partial_\zeta N_m) & (y_m \partial_\zeta N_m) & (z_m \partial_\zeta N_m) \end{bmatrix}$$

Using the transformation we obtain

$$\nabla_{(\xi, \eta, \zeta)} N = \begin{Bmatrix} \partial_\xi N \\ \partial_\eta N \\ \partial_\zeta N \end{Bmatrix} = J \cdot \nabla N \qquad (16\text{-}20)$$

$$\nabla N = J^{-1} \cdot \nabla_{(\xi, \eta, \zeta)} N$$

for the derivatives and

$$e_{(\xi, \eta, \zeta)} = \begin{Bmatrix} e_\xi \\ e_\eta \\ e_\zeta \end{Bmatrix} = J \cdot e \qquad e = J^{-1} \cdot e_{(\xi, \eta, \zeta)} \quad (16\text{-}21)$$

for the gravity vector, where J^{-1} is the inverse Jacobian. Using these relationships the equivalent formulation of the Darcy velocity (16-2) in local coordinates is given by

$$v = -Kf_\mu \cdot J^{-1} \cdot (\nabla_{(\xi, \eta, \zeta)} h + \tilde{\rho} J \cdot e) \qquad (16\text{-}22a)$$

or

$$v = -Kf_\mu \cdot J^{-1} \cdot (\nabla_{(\xi, \eta, \zeta)} h + \tilde{\rho} e_{(\xi, \eta, \zeta)}) \qquad (16\text{-}22b)$$

16.4.1.2 The new formulation

Introducing the following integral functions

$$\left. \begin{aligned} H_\xi &= H_\xi(\xi, \eta, \zeta) = \int_0^\xi \tilde{\rho}(\theta, \eta, \zeta) e_\xi(\theta, \eta, \zeta) d\theta \\ H_\eta &= H_\eta(\xi, \eta, \zeta) = \int_0^\eta \tilde{\rho}(\xi, \theta, \zeta) e_\eta(\xi, \theta, \zeta) d\theta \\ H_\zeta &= H_\zeta(\xi, \eta, \zeta) = \int_0^\zeta \tilde{\rho}(\xi, \eta, \theta) e_\zeta(\xi, \eta, \theta) d\theta \end{aligned} \right\} \quad (16\text{-}23)$$

Since

$$\begin{Bmatrix} \partial_\xi H_\xi \\ \partial_\eta H_\eta \\ \partial_\zeta H_\zeta \end{Bmatrix} = \tilde{\rho} e_{(\xi, \eta, \zeta)} \qquad (16\text{-}24)$$

we can write the Darcy velocity (16-22b) in an equivalent form

$$v = -Kf_\mu \cdot J^{-1} \cdot \begin{Bmatrix} \partial_\xi (h + H_\xi) \\ \partial_\eta (h + H_\eta) \\ \partial_\zeta (h + H_\zeta) \end{Bmatrix} \qquad (16\text{-}25)$$

These new integral functions H_ξ, H_η, H_ζ allows us to obtain the same spatial variability for both the head (h)-term and the gravity term.

The consistency of (16-25) in the definition of (16-7) can be proved. Assuming the gravity acts in the z-direction, i.e., $\tilde{\rho}(x, y, z) = \tilde{\rho}(x_o, y_o, z)$ we can write

$$H_\xi = \int_0^\xi \tilde{\rho}(x_o, y_o, z(\theta, \eta, \zeta)) e_\xi d\theta \qquad (16\text{-}26)$$

$$= e_z \int_{z_0}^{z(\xi, \eta, \zeta)} \tilde{\rho}(x_o, y_o, \theta) d\theta$$

where $(x_0, y_0, z_0) = (x(0, 0, 0), y(0, 0, 0), z(0, 0, 0))$ and similarly for H_η and H_ζ.

In the finite element method the functions $h, H_\xi, H_\eta, H_\zeta$ are interpolated by their nodal basis functions:

$$\left. \begin{array}{l} h = \sum_m h_m N_m(\xi, \eta, \zeta) \\[1em] H_\xi = \sum_m H_{\xi m} N_m(\xi, \eta, \zeta) \\[1em] H_\eta = \sum_m H_{\eta m} N_m(\xi, \eta, \zeta) \\[1em] H_\zeta = \sum_m H_{\zeta m} N_m(\xi, \eta, \zeta) \end{array} \right\} \qquad (16\text{-}27)$$

and we obtain the velocity (16-25) in the discretized formulation

$$\boldsymbol{v} = -\boldsymbol{K} f_\mu \cdot \boldsymbol{J}^{-1} \cdot \sum_m \left\{ \begin{array}{l} (h_m + H_{\xi m}) \partial_\xi N_m(\xi, \eta, \zeta) \\[0.5em] (h_m + H_{\eta m}) \partial_\eta N_m(\xi, \eta, \zeta) \\[0.5em] (h_m + H_{\zeta m}) \partial_\zeta N_m(\xi, \eta, \zeta) \end{array} \right\} \qquad (16\text{-}28)$$

which represents a fully consistent approximation of the Darcy velocities. We solve (16-28) for given heads h and the values of the H_ξ, H_η, H_ζ-functions at the nodes m. The nodal quantities $H_{\xi m}, H_{\eta m}, H_{\zeta m}$ are dependent on the finite element types and will be evaluated next for *linear* elements in two and three dimensions. In doing this, the relative density $\tilde{\rho}$ in the gravity term is interpolated according to

$$\tilde{\rho} = \sum_m \tilde{\rho}_m N_m(\xi, \eta, \zeta) \qquad (16\text{-}29)$$

where $\tilde{\rho}_m$ are the density values at the node m. In FEFLOW the following constitutive relationship is used for the density ρ

$$\rho = \rho_0 \left[1 + \frac{\alpha}{C_s - C_o}(C - C_o) - \beta(T - T_o) \right] \qquad (16\text{-}30)$$

as a function of the concentration C and the temperature T. Accordingly, the relative density $\tilde{\rho}$ (16-3) is given by

$$\tilde{\rho} = \frac{\alpha}{C_s - C_o}(C - C_o) - \beta(T - T_o) \qquad (16\text{-}31)$$

and the finite element expansion (16-29) can alternatively be written as

$$\tilde{\rho} = \frac{\alpha}{C_s - C_o} \sum_m N_m(C_m - C_o) - \beta \sum_m N_m(T_m - T_o) \qquad (16\text{-}32)$$

or

$$\tilde{\rho}_m = \frac{\alpha}{C_s - C_o}(C_m - C_o) - \beta(T_m - T_o) \qquad (16\text{-}33)$$

in relation to the expansion (16-29).

16.4.1.3 The nodal quantities $H_{\xi m}, H_{\eta m}, H_{\zeta m}$ of the integral functions

Linear triangular element in two dimensions

For a triangle we use the local coordinates as indicated in Fig. 16.3.
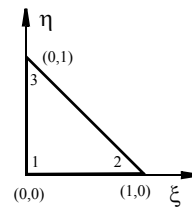
$$N_1 = 1 - \xi - \eta \qquad \partial_\xi N_1 = -1 \qquad \partial_\eta N_1 = -1$$
$$N_2 = \xi \qquad \partial_\xi N_2 = 1 \qquad \partial_\eta N_2 = 0$$
$$N_3 = \eta \qquad \partial_\xi N_3 = 0 \qquad \partial_\eta N_3 = 1$$

Figure 16.3 Local coordinates, shape functions and local derivatives for the linear triangular element.

The Jacobian $\boldsymbol{J}$ (16-19) appears independent of the local coordinates (ξ, η) and the gravity (e_ξ, e_η) in the local coordinates from (16-21) is a constant vector. Accordingly, we can write

$$
\begin{aligned}
H_\xi &= \int_0^\xi \tilde{\rho}(\theta, \eta) e_\xi(\theta, \eta) d\theta = e_\xi \int_0^\xi \tilde{\rho}(\theta, \eta) d\theta \\
&= e_\xi \int_0^\xi \left(\sum_{m=1}^3 N_m(\theta, \eta) \tilde{\rho}_m \right) d\theta \\
&= e_\xi \int_0^\xi [(1 - \theta - \eta)\tilde{\rho}_1 + \theta\tilde{\rho}_2 + \eta\tilde{\rho}_3] d\theta \\
&= e_\xi \left[\left(\xi - \frac{\xi^2}{2} - \xi\eta \right) \tilde{\rho}_1 + \frac{\xi^2}{2}\tilde{\rho}_2 + \xi\eta\tilde{\rho}_3 \right]
\end{aligned}
\qquad (16\text{-}34)
$$

and similarly for H_η. From the integrals we find the nodal values for H_ξ and H_η as

$$
\left.
\begin{aligned}
H_\xi(0, 0) &= H_{\xi 1} = 0 \\
H_\xi(1, 0) &= H_{\xi 2} = \frac{1}{2} e_\xi (\tilde{\rho}_1 + \tilde{\rho}_2) \\
H_\xi(0, 1) &= H_{\xi 3} = 0
\end{aligned}
\right\}
\qquad (16\text{-}35a)
$$

$$
\left.
\begin{aligned}
H_\eta(0, 0) &= H_{\eta 1} = 0 \\
H_\eta(1, 0) &= H_{\eta 2} = 0 \\
H_\eta(0, 1) &= H_{\eta 3} = \frac{1}{2} e_\eta (\tilde{\rho}_1 + \tilde{\rho}_3)
\end{aligned}
\right\}
\qquad (16\text{-}35b)
$$

Now we can express the gravity term (16-24) in local coordinates as

$$\tilde{\rho}\boldsymbol{e}_{(\xi,\eta)} = \left\{ \begin{matrix} \partial_\xi H_\xi \\ \partial_\eta H_\eta \end{matrix} \right\}$$

representing a consistent approximation in which the density is appropriately averaged in the gravitational directions.

$$= \left\{ \begin{matrix} \partial_\xi N_1 H_{\xi 1} + \partial_\xi N_2 H_{\xi 2} + \partial_\xi N_3 H_{\xi 3} \\ \partial_\eta N_1 H_{\eta 1} + \partial_\eta N_2 H_{\eta 2} + \partial_\eta N_3 H_{\eta 3} \end{matrix} \right\} \quad (16\text{-}36)$$

$$= \frac{1}{2} \left\{ \begin{matrix} e_\xi(\tilde{\rho}_1 + \tilde{\rho}_2) \\ e_\eta(\tilde{\rho}_1 + \tilde{\rho}_3) \end{matrix} \right\}$$

$$N_1 = \frac{1}{4}(1-\xi)(1-\eta) \qquad \partial_\xi N_1 = -\frac{1}{4}(1-\eta) \qquad \partial_\eta N_1 = -\frac{1}{4}(1-\xi)$$

$$N_2 = \frac{1}{4}(1+\xi)(1-\eta) \qquad \partial_\xi N_2 = \frac{1}{4}(1-\eta) \qquad \partial_\eta N_2 = -\frac{1}{4}(1+\xi)$$

$$N_3 = \frac{1}{4}(1+\xi)(1+\eta) \qquad \partial_\xi N_3 = \frac{1}{4}(1+\eta) \qquad \partial_\eta N_3 = \frac{1}{4}(1+\xi)$$

$$N_4 = \frac{1}{4}(1-\xi)(1+\eta) \qquad \partial_\xi N_4 = -\frac{1}{4}(1+\eta) \qquad \partial_\eta N_4 = \frac{1}{4}(1-\xi)$$
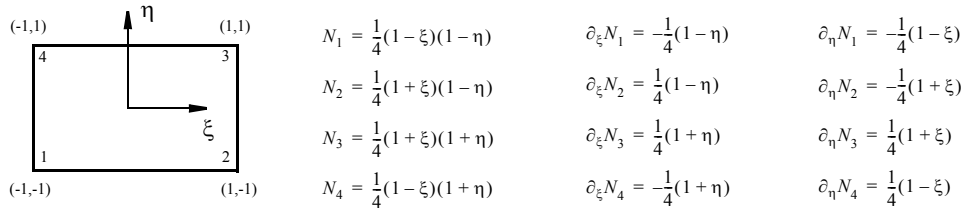
Figure 16.4 Local coordinates, shape functions and local derivatives for the linear quadrilateral element.

Linear quadrilateral element in two dimensions
The local coordinates, the related shape functions and the local derivatives are shown in Fig. 16.4 for the linear quadrilateral element. While for this element the Jacobian $\boldsymbol{J}$ (16-19) is in general space-dependent, the gravity vector (16-21) in local coordinates takes the special form

$$\left\{ \begin{matrix} e_\xi \\ e_\eta \end{matrix} \right\} = \left\{ \begin{matrix} e_\xi(\eta) \\ e_\eta(\xi) \end{matrix} \right\} \quad (16\text{-}37)$$

Similarly to the above triangular element, we can compute the integral functions H_ξ, H_η at the corner nodes

m for the linear quadrilateral element as

$$H_\xi(-1,-1) = H_{\xi 1} = -\frac{1}{4}e_\xi(-1)(3\tilde{\rho}_1 + \tilde{\rho}_2)$$

$$H_\xi(1,-1) = H_{\xi 2} = \frac{1}{4}e_\xi(-1)(\tilde{\rho}_1 + 3\tilde{\rho}_2)$$

$$H_\xi(1,1) = H_{\xi 3} = \frac{1}{4}e_\xi(1)(3\tilde{\rho}_3 + \tilde{\rho}_4)$$

$$H_\xi(-1,1) = H_{\xi 4} = -\frac{1}{4}e_\xi(1)(\tilde{\rho}_3 + 3\tilde{\rho}_4)$$

$$(16\text{-}38a)$$

$$H_\eta(-1,-1) = H_{\eta 1} = -\frac{1}{4}e_\eta(-1)(3\tilde{\rho}_1 + \tilde{\rho}_4)$$

$$H_\eta(1,-1) = H_{\eta 2} = -\frac{1}{4}e_\eta(1)(3\tilde{\rho}_2 + \tilde{\rho}_3)$$

$$H_\eta(1,1) = H_{\eta 3} = \frac{1}{4}e_\eta(1)(\tilde{\rho}_2 + 3\tilde{\rho}_3)$$

$$H_\eta(-1,1) = H_{\eta 4} = \frac{1}{4}e_\eta(-1)(\tilde{\rho}_1 + 3\tilde{\rho}_4)$$

(16-38b)

For the linear quadrilateral element the consistent approximation (16-39) can be in recognized as the consistent formulation previously introduced by Voss[20], where the gravity term is averaged in a directional manner, so for instance

$$\tilde{\rho}\boldsymbol{e}_{(\xi = 1, \eta = 1)} = \frac{1}{2}\left\{ \begin{matrix} e_\xi(1)(\tilde{\rho}_3 + \tilde{\rho}_4) \\ e_\eta(1)(\tilde{\rho}_2 + \tilde{\rho}_3) \end{matrix} \right\}$$

(16-40)

The gravity term (16-24) written in local coordinates yields

$$\tilde{\rho}\boldsymbol{e}_{(\xi,\eta)} = \left\{ \begin{matrix} \partial_\xi H_\xi \\ \partial_\eta H_\eta \end{matrix} \right\}$$

$$= \left\{ \begin{matrix} \partial_\xi N_1 H_{\xi 1} + \partial_\xi N_2 H_{\xi 2} + \partial_\xi N_3 H_{\xi 3} + \partial_\xi N_4 H_{\xi 4} \\ \partial_\eta N_1 H_{\eta 1} + \partial_\eta N_2 H_{\eta 2} + \partial_\eta N_3 H_{\eta 3} + \partial_\eta N_4 H_{\eta 4} \end{matrix} \right\}$$

(16-39)

$$= \frac{1}{4}\left\{ \begin{matrix} e_\xi(-1)(1-\eta)(\tilde{\rho}_1 + \tilde{\rho}_2) + e_\xi(1)(1+\eta)(\tilde{\rho}_3 + \tilde{\rho}_4) \\ e_\eta(-1)(1-\xi)(\tilde{\rho}_1 + \tilde{\rho}_4) + e_\eta(1)(1+\xi)(\tilde{\rho}_2 + \tilde{\rho}_3) \end{matrix} \right\}$$

$$N_1 = \frac{1}{2}(1 - \xi - \eta)(1 + \zeta)$$

$$N_2 = \frac{1}{2}\xi(1 + \zeta)$$

$$N_3 = \frac{1}{2}\eta(1 + \zeta)$$

$$N_4 = \frac{1}{2}(1 - \xi - \eta)(1 - \zeta)$$

$$N_5 = \frac{1}{2}\xi(1 - \zeta)$$

$$N_6 = \frac{1}{2}\eta(1 - \zeta)$$

$$\partial_\xi N_1 = -\frac{1}{2}(1 + \zeta)$$

$$\partial_\xi N_2 = \frac{1}{2}(1 + \zeta)$$

$$\partial_\xi N_3 = 0$$

$$\partial_\xi N_4 = -\frac{1}{2}(1 - \zeta)$$

$$\partial_\xi N_5 = \frac{1}{2}(1 - \zeta)$$

$$\partial_\xi N_6 = 0$$

$$\partial_\eta N_1 = -\frac{1}{2}(1 + \zeta)$$

$$\partial_\eta N_2 = 0$$

$$\partial_\eta N_3 = \frac{1}{2}(1 + \zeta)$$

$$\partial_\eta N_4 = -\frac{1}{2}(1 - \zeta)$$

$$\partial_\eta N_5 = 0$$

$$\partial_\eta N_6 = \frac{1}{2}(1 - \zeta)$$

$$\partial_\zeta N_1 = \frac{1}{2}(1 - \xi - \eta)$$

$$\partial_\zeta N_2 = \frac{1}{2}\xi$$

$$\partial_\zeta N_3 = \frac{1}{2}\eta$$

$$\partial_\zeta N_4 = -\frac{1}{2}(1 - \xi - \eta)$$

$$\partial_\zeta N_5 = -\frac{1}{2}\xi$$

$$\partial_\zeta N_6 = -\frac{1}{2}\eta$$
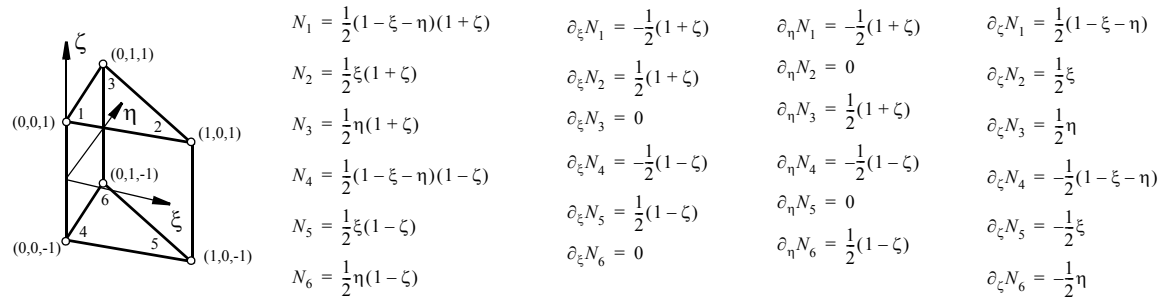
Figure 16.5 Local coordinates, shape functions and local derivatives for the linear pentahedral element.

the gravity vector (16-21) is in local coordinates

Linear pentahedral (triangular prismatic) element in three dimensions

The pentahedral element and its local functions are shown in Fig. 16.5. Specifying the Jacobian $\boldsymbol{J}$ (16-19)

$$\begin{Bmatrix} e_\xi \\ e_\eta \\ e_\zeta \end{Bmatrix} = \begin{Bmatrix} e_\xi(\zeta) \\ e_\eta(\zeta) \\ e_\zeta(\xi, \eta) \end{Bmatrix} \qquad (16\text{-}41)$$

The integral functions H_ξ, H_η, H_ζ at the corner nodes m for the linear pentahedral element are then

$$\left.\begin{aligned}
H_\xi(0, 0, 1) &= H_{\xi 1} = 0 \\
H_\xi(1, 0, 1) &= H_{\xi 2} = \frac{1}{2}e_\xi(1)(\tilde{\rho}_1 + \tilde{\rho}_2) \\
H_\xi(0, 1, 1) &= H_{\xi 3} = 0 \\
H_\xi(0, 0, -1) &= H_{\xi 4} = 0 \\
H_\xi(1, 0, -1) &= H_{\xi 5} = \frac{1}{2}e_\xi(-1)(\tilde{\rho}_4 + \tilde{\rho}_5) \\
H_\xi(0, 1, -1) &= H_{\xi 6} = 0
\end{aligned}\right\} \qquad (16\text{-}42a)$$

$$\left.\begin{aligned}
H_\eta(0, 0, 1) &= H_{\eta 1} = 0 \\
H_\eta(1, 0, 1) &= H_{\eta 2} = 0 \\
H_\eta(0, 1, 1) &= H_{\eta 3} = \frac{1}{2}e_\eta(1)(\tilde{\rho}_1 + \tilde{\rho}_3) \\
H_\eta(0, 0, -1) &= H_{\eta 4} = 0 \\
H_\eta(1, 0, -1) &= H_{\eta 5} = 0 \\
H_\eta(0, 1, -1) &= H_{\eta 6} = \frac{1}{2}e_\eta(-1)(\tilde{\rho}_4 + \tilde{\rho}_6)
\end{aligned}\right\} \qquad (16\text{-}42b)$$

$$\left.\begin{aligned}
H_\zeta(0, 0, 1) &= H_{\zeta 1} = \frac{1}{4}e_\zeta(0, 0)(3\tilde{\rho}_1 + \tilde{\rho}_4) \\
H_\zeta(1, 0, 1) &= H_{\zeta 2} = \frac{1}{4}e_\zeta(1, 0)(3\tilde{\rho}_2 + \tilde{\rho}_5) \\
H_\zeta(0, 1, 1) &= H_{\zeta 3} = \frac{1}{4}e_\zeta(0, 1)(3\tilde{\rho}_3 + \tilde{\rho}_6) \\
H_\zeta(0, 0, -1) &= H_{\zeta 4} = -\frac{1}{4}e_\zeta(0, 0)(\tilde{\rho}_1 + 3\tilde{\rho}_4) \\
H_\zeta(1, 0, -1) &= H_{\zeta 5} = -\frac{1}{4}e_\zeta(1, 0)(\tilde{\rho}_2 + 3\tilde{\rho}_5) \\
H_\zeta(0, 1, -1) &= H_{\zeta 6} = -\frac{1}{4}e_\zeta(0, 1)(\tilde{\rho}_3 + 3\tilde{\rho}_6)
\end{aligned}\right\} \qquad (16\text{-}42c)$$
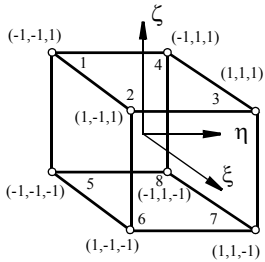
$$N_1 = \frac{1}{8}(1-\xi)(1-\eta)(1+\zeta) \qquad \partial_\xi N_1 = -\frac{1}{8}(1-\eta)(1+\zeta) \qquad \partial_\eta N_1 = -\frac{1}{8}(1-\xi)(1+\zeta) \qquad \partial_\zeta N_1 = \frac{1}{8}(1-\xi)(1-\eta)$$

$$N_2 = \frac{1}{8}(1+\xi)(1-\eta)(1+\zeta) \qquad \partial_\xi N_2 = \frac{1}{8}(1-\eta)(1+\zeta) \qquad \partial_\eta N_2 = -\frac{1}{8}(1+\xi)(1+\zeta) \qquad \partial_\zeta N_2 = \frac{1}{8}(1+\xi)(1-\eta)$$

$$N_3 = \frac{1}{8}(1+\xi)(1+\eta)(1+\zeta) \qquad \partial_\xi N_3 = \frac{1}{8}(1+\eta)(1+\zeta) \qquad \partial_\eta N_3 = \frac{1}{8}(1+\xi)(1+\zeta) \qquad \partial_\zeta N_3 = \frac{1}{8}(1+\xi)(1+\eta)$$

$$N_4 = \frac{1}{8}(1-\xi)(1+\eta)(1+\zeta) \qquad \partial_\xi N_4 = -\frac{1}{8}(1+\eta)(1+\zeta) \qquad \partial_\eta N_4 = \frac{1}{8}(1-\xi)(1+\zeta) \qquad \partial_\zeta N_4 = \frac{1}{8}(1-\xi)(1+\eta)$$

$$N_5 = \frac{1}{8}(1-\xi)(1-\eta)(1-\zeta) \qquad \partial_\xi N_5 = -\frac{1}{8}(1-\eta)(1-\zeta) \qquad \partial_\eta N_5 = -\frac{1}{8}(1-\xi)(1-\zeta) \qquad \partial_\zeta N_5 = -\frac{1}{8}(1-\xi)(1-\eta)$$

$$N_6 = \frac{1}{8}(1+\xi)(1-\eta)(1-\zeta) \qquad \partial_\xi N_6 = \frac{1}{8}(1-\eta)(1-\zeta) \qquad \partial_\eta N_6 = -\frac{1}{8}(1+\xi)(1-\zeta) \qquad \partial_\zeta N_6 = -\frac{1}{8}(1+\xi)(1-\eta)$$

$$N_7 = \frac{1}{8}(1+\xi)(1+\eta)(1-\zeta) \qquad \partial_\xi N_7 = \frac{1}{8}(1+\eta)(1-\zeta) \qquad \partial_\eta N_7 = \frac{1}{8}(1+\xi)(1-\zeta) \qquad \partial_\zeta N_7 = -\frac{1}{8}(1+\xi)(1+\eta)$$

$$N_8 = \frac{1}{8}(1-\xi)(1+\eta)(1-\zeta) \qquad \partial_\xi N_8 = -\frac{1}{8}(1+\eta)(1-\zeta) \qquad \partial_\eta N_8 = \frac{1}{8}(1-\xi)(1-\zeta) \qquad \partial_\zeta N_8 = -\frac{1}{8}(1-\xi)(1+\eta)$$

Figure 16.6 Local coordinates, shape functions and local derivatives for the linear hexahedral element.

Linear hexahedral (brick) element in three dimensions

The hexahedral element and its local functions are displayed in Fig. 16.6.

The gravity vector (16-21) for this element is in local coordinates

$$\left\{\begin{array}{c} e_\xi \\ e_\eta \\ e_\zeta \end{array}\right\} = \left\{\begin{array}{c} e_\xi(\eta, \zeta) \\ e_\eta(\xi, \zeta) \\ e_\zeta(\xi, \eta) \end{array}\right\} \qquad (16\text{-}43)$$

The integral functions H_ξ, H_η, H_ζ at the corner nodes m for the linear hexahedral element can be derived as

$$\left.\begin{array}{l} H_\xi(-1,-1,1) = H_{\xi 1} = -\frac{1}{4}e_\xi(-1,1)(3\tilde{\rho}_1 + \tilde{\rho}_2) \\[6pt] H_\xi(1,-1,1) = H_{\xi 2} = \frac{1}{4}e_\xi(-1,1)(\tilde{\rho}_1 + 3\tilde{\rho}_2) \\[6pt] H_\xi(1,1,1) = H_{\xi 3} = \frac{1}{4}e_\xi(1,1)(3\tilde{\rho}_3 + \tilde{\rho}_4) \\[6pt] H_\xi(-1,1,1) = H_{\xi 4} = -\frac{1}{4}e_\xi(1,1)(\tilde{\rho}_3 + 3\tilde{\rho}_4) \\[6pt] H_\xi(-1,-1,-1) = H_{\xi 5} = -\frac{1}{4}e_\xi(-1,-1)(3\tilde{\rho}_5 + \tilde{\rho}_6) \\[6pt] H_\xi(1,-1,-1) = H_{\xi 6} = \frac{1}{4}e_\xi(-1,-1)(\tilde{\rho}_5 + 3\tilde{\rho}_6) \\[6pt] H_\xi(1,1,-1) = H_{\xi 7} = \frac{1}{4}e_\xi(1,-1)(3\tilde{\rho}_7 + \tilde{\rho}_8) \\[6pt] H_\xi(-1,1,-1) = H_{\xi 8} = -\frac{1}{4}e_\xi(1,-1)(\tilde{\rho}_7 + 3\tilde{\rho}_8) \end{array}\right\} \qquad (16\text{-}44a)$$

$$\left.\begin{array}{l} H_\eta(-1,-1,1) = H_{\eta 1} = -\frac{1}{4}e_\eta(-1,1)(3\tilde{\rho}_1 + \tilde{\rho}_4) \\[6pt] H_\eta(1,-1,1) = H_{\eta 2} = -\frac{1}{4}e_\eta(1,1)(3\tilde{\rho}_2 + \tilde{\rho}_3) \\[6pt] H_\eta(1,1,1) = H_{\eta 3} = \frac{1}{4}e_\eta(1,1)(\tilde{\rho}_2 + 3\tilde{\rho}_3) \\[6pt] H_\eta(-1,1,1) = H_{\eta 4} = \frac{1}{4}e_\eta(-1,1)(\tilde{\rho}_1 + 3\tilde{\rho}_4) \\[6pt] H_\eta(-1,-1,-1) = H_{\eta 5} = -\frac{1}{4}e_\eta(-1,-1)(3\tilde{\rho}_5 + \tilde{\rho}_8) \\[6pt] H_\eta(1,-1,-1) = H_{\eta 6} = -\frac{1}{4}e_\eta(1,-1)(3\tilde{\rho}_6 + \tilde{\rho}_7) \\[6pt] H_\eta(1,1,-1) = H_{\eta 7} = \frac{1}{4}e_\eta(1,-1)(\tilde{\rho}_6 + 3\tilde{\rho}_7) \\[6pt] H_\eta(-1,1,-1) = H_{\eta 8} = \frac{1}{4}e_\eta(-1,-1)(\tilde{\rho}_5 + 3\tilde{\rho}_8) \end{array}\right\} \qquad (16\text{-}44b)$$

$$\left.\begin{array}{l} H_\zeta(-1,-1,1) = H_{\zeta 1} = \frac{1}{4}e_\zeta(-1,-1)(3\tilde{\rho}_1 + \tilde{\rho}_5) \\[6pt] H_\zeta(1,-1,1) = H_{\zeta 2} = \frac{1}{4}e_\zeta(1,-1)(3\tilde{\rho}_2 + \tilde{\rho}_6) \\[6pt] H_\zeta(1,1,1) = H_{\zeta 3} = \frac{1}{4}e_\zeta(1,1)(3\tilde{\rho}_3 + \tilde{\rho}_7) \\[6pt] H_\zeta(-1,1,1) = H_{\zeta 4} = \frac{1}{4}e_\zeta(-1,1)(3\tilde{\rho}_4 + \tilde{\rho}_8) \\[6pt] H_\zeta(-1,-1,-1) = H_{\zeta 5} = -\frac{1}{4}e_\zeta(-1,-1)(\tilde{\rho}_1 + 3\tilde{\rho}_5) \\[6pt] H_\zeta(1,-1,-1) = H_{\zeta 6} = -\frac{1}{4}e_\zeta(1,-1)(\tilde{\rho}_2 + 3\tilde{\rho}_6) \\[6pt] H_\zeta(1,1,-1) = H_{\zeta 7} = -\frac{1}{4}e_\zeta(1,1)(\tilde{\rho}_3 + 3\tilde{\rho}_7) \\[6pt] H_\zeta(-1,1,-1) = H_{\zeta 8} = -\frac{1}{4}e_\zeta(-1,1)(\tilde{\rho}_4 + 3\tilde{\rho}_8) \end{array}\right\} \qquad (16\text{-}44c)$$

For the hexahedral (brick) element the consistent formulation of the gravity term in form of the integral functions H_ξ, H_η, H_ζ (*cf.* (16-44a), (16-44b) and (16-44c), respectively), is equivalent to the formulation given by Leijnse[17]. This should be exemplified for the $\tilde{\rho}e_\xi$-component of the gravity term:

$$\tilde{\rho}e_\xi = \sum_{m=1}^{8} \partial_\xi N_m H_{\xi m} = \frac{1}{8}[e_\xi(-1,1)(1-\eta)(1+\zeta)(\tilde{\rho}_1+\tilde{\rho}_2) + \\ + e_\xi(1,1)(1+\eta)(1+\zeta)(\tilde{\rho}_3+\tilde{\rho}_4) + \\ + e_\xi(-1,-1)(1-\eta)(1-\zeta)(\tilde{\rho}_5+\tilde{\rho}_6) + \\ + e_\xi(1,-1)(1+\eta)(1-\zeta)(\tilde{\rho}_7+\tilde{\rho}_8)] \quad (16\text{-}45)$$

16.4.2 Continuous consistent velocities

The computation of the consistent velocities (16-28) is performed elementwise in a standard manner, i.e.,

$$v_p^e = -K^e f_\mu^e \cdot J^{-1} \cdot \sum_m \begin{Bmatrix} (h_m + H_{\xi m})\partial_\xi N_m(\xi_p^e, \eta_p^e, \zeta_p^e) \\ (h_m + H_{\eta m})\partial_\eta N_m(\xi_p^e, \eta_p^e, \zeta_p^e) \\ (h_m + H_{\zeta m})\partial_\zeta N_m(\xi_p^e, \eta_p^e, \zeta_p^e) \end{Bmatrix} (16\text{-}46)$$

by using (16-21)

$$e_{(\xi,\eta,\zeta)} = J \cdot e \quad (16\text{-}47)$$

for the local gravity component in evaluating the nodal integral functions $H_{\xi m}, H_{\eta m}, H_{\zeta m}$ according to (16-35a)-(16-35b), (16-38a)-(16-38b), (16-42a)-(16-42c) and (16-44a)-(16-44c). In (16-47) and (16-46) the Jacobian J and the global derivatives $\nabla(.) = J^{-1} \cdot \nabla_{(\xi,\eta,\zeta)}(.)$ are evaluated at the Gauss points p for each element e.

The element-by-element technique (16-46) leads naturally to a consistent velocity field, which is in general discontinuous at the nodes m. To obtain continu-ous velocities a local smoothing technique such as described in the Appendix B can be easily applied. Obviously, the smoothing procedure has no effect on the consistency of the velocity. Since the velocities $v_{p \to m}^e$ for each element e are always consistent at the node m an element-patch-averaged velocity $v_m = \sum_e v_m^e / n_p$ must be consistent too.

16.5 Examples

16.5.1 Hydrostatic condition in a closed porous box

Let us consider a rectangular closed domain as shown in Fig. 16.7. At initial time a stable saltwater layer with a salinity (concentration) of $C = C_s$ exists below freshwater with $C = C_0 = 0$ separated by a horizontal sharp interface in the middle of the domain. The domain is impervious with respect to both the flow and the mass transport. The fluid density contrast α defined by

$$\alpha = \frac{\rho_s - \rho_o}{\rho_o} \quad \text{with} \quad \rho_s = \rho(C_s), \quad \rho_o = \rho(C_o) \quad (16\text{-}48)$$

amounts to a value of 0.03.

16. Consistent velocity approximation in the finite-element simulation of density-dependent mass and heat transport processes
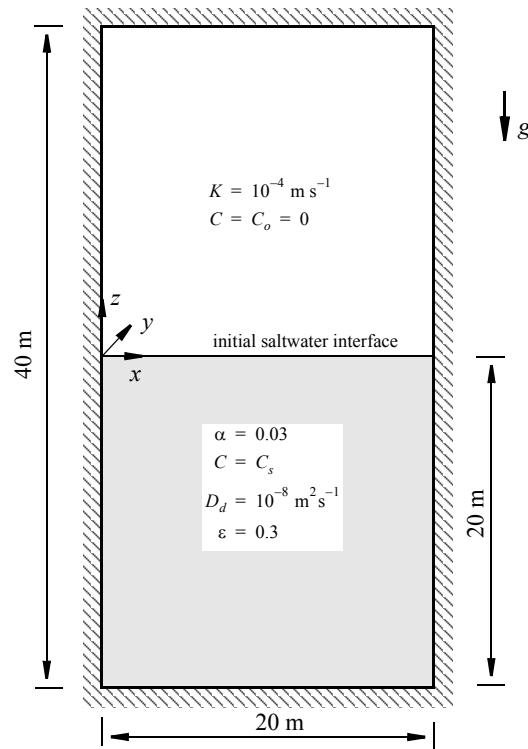
Figure 16.7 Cross-sectional view of the initially stratified saltwater below freshwater problem in a closed porous box.

The problem is hydrostatic over all times and the fluid motion within the box should be zero or, in the numerical sense, negligibly small. Due to the molecular diffusion D_d the saltwater mixes and the initially sharp saltwater interface (narrow transition zone) spreads in time. This process must be independent of the density effect. Accordingly, we have to compare the results of the saltwater interface spreading for the case without density coupling against the cases, where density effects are included. As a reference solution we compute the problem for $\alpha \equiv 0$ based on a fine temporal and spatial discretization.

We simulated the density-dependent problem for quadrilateral and triangular meshes both in two and three dimensions by using the different velocity approximations. The findings are practically the same to that depicted in Fig. 16.8 for the two-dimensional quadrilateral elements.

The new formulation for the consistent velocity approximation by the Frolkovic-Knabner method agrees very well with the reference solution. We tested both without dispersion ($\beta_L = \beta_T = 0$) and with dispersion effects ($\beta_L = 5\text{m}, \beta_T = 0.5\text{m}$). The results are identical. In contrast, the old formulation which is based on a locally smoothed nonconsistent velocity approximation gives erroneous results in form of a smeared density profile depart from the reference solution. Expectedly, this effects increases if dispersion ($\beta_L = 5\text{m}, \beta_T = 0.5\text{m}$) is taken into account.

It is obvious the old formulation computes spurious local velocities at the interface nodes which lead to an artificially increased spreading of the salinity which has an effect similar to numerical dispersion. It becomes clear if we magnify the local velocities of the interface node at beginning of the simulation as illustrated in Fig. 16.9. The old formulation gives spurious velocities at the local points of the interface in the order of about $10^{-2}\,\text{md}^{-1}$, while outside the interface the nodal velocities are very small in order of only $10^{-7}\,\text{md}^{-1}$. In contrast, the velocities for the new for-

mulation remain generally small in all points of the domain in the order of about $10^{-7}\,md^{-1}$, which represents the numerical noise in the velocity field for this type of problem.
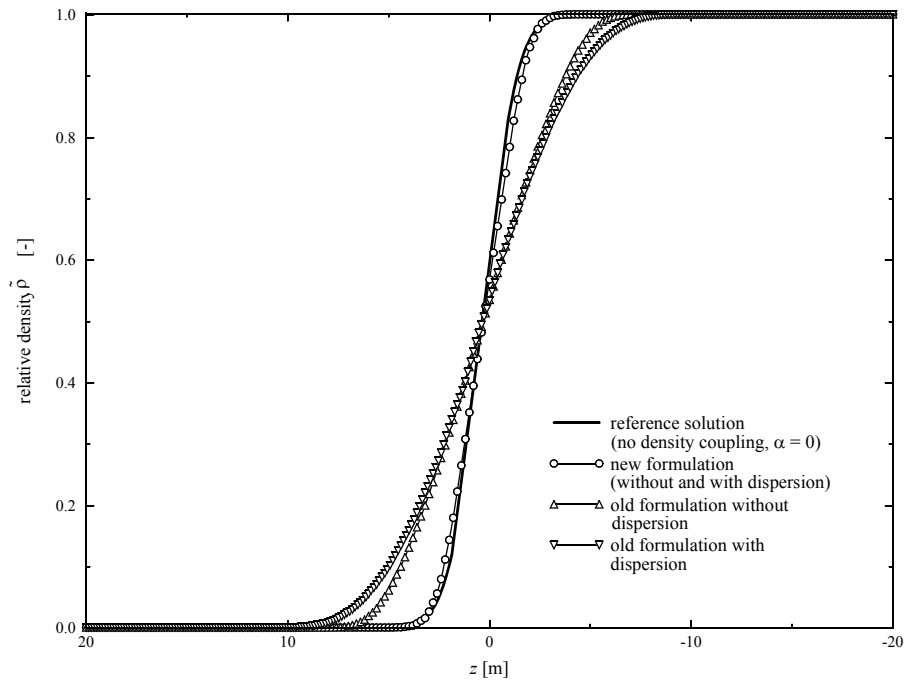
Figure 16.8 Computed density profiles $\tilde{\rho}(x, z)$, $x = 10m$, $20m \leq z \leq -20m$ at time $t = 10^3$ days for different solutions using quadrilateral elements: Reference solution is obtained without density effects for a fine vertical mesh; the other solutions are simulated on an uniform 32x64 mesh of quadrilateral elements: new formulation represents the Frolkovic-Knabner algorithm for consistent velocities combined with local smoothing, old formulation is the local smoothing of the basically nonconsistent velocities.

old formulation   new formulation
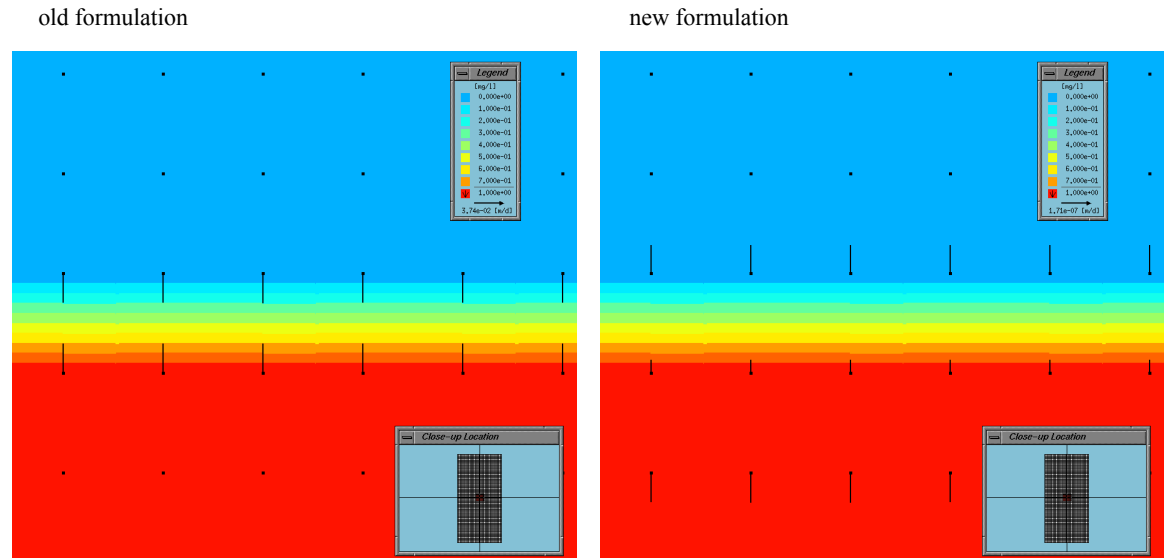
Figure 16.9 Local velocities at the saltwater interface computed by the old versus the new formulation.

16.5.2 The Elder problem revisited

The Elder problem describes a free convection process in a vertical cross-section which is extensively studied in past by various authors in using different numerical methods[1,5,6,10,11,21]. It is defined in Fig. 16.10 with respect to a saline problem type so as basically proposed by Voss and Souza[21].
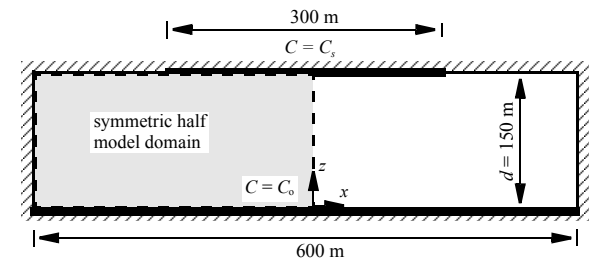
Figure 16.10 Definition of the two-dimensional Elder problem and the used symmetric half model domain.

The Elder problem is studied for a (solutal) Rayleigh

number $Ra = 400$, where Ra is defined as

$$Ra = \frac{\alpha \cdot K \cdot d}{\varepsilon \cdot D_d} \qquad (16\text{-}49)$$

(for more details see for instance[6,14]).

Recently, Frolkovic and De Schepper[10] presented new results for the Elder problem. They achieved grid convergence by a systematical refinement of the mesh for the symmetric half of the domain (Fig. 16.10) using a *grid level l* in the range of 4 to 8. For an uniform discretization by quadrilateral elements the number of elements of the half domain is given by

$$ne = 2 \cdot 4^l \qquad (16\text{-}50)$$

Frolkovic and De Schepper's findings have been confirmed by FEFLOW computations, however, based on the old formulation of the velocity approximation, where grid levels between 4 and 9 were applied. Now, it is interesting to see the effect of the new formulation of the consistent velocity approximation on these results. A comparison between the two formulations is shown in Fig. 16.11 for the Elder problem at a grid level $l = 7$. As seen there are only slight differences in the salinity and streamline patterns which have practically no effect on the history of the cellular convection process. This means the numerical quality of the velocity field for the old and the new formulation is efficiently the same for a problem where the density contrast (say Rayleigh number Ra) is moderate.
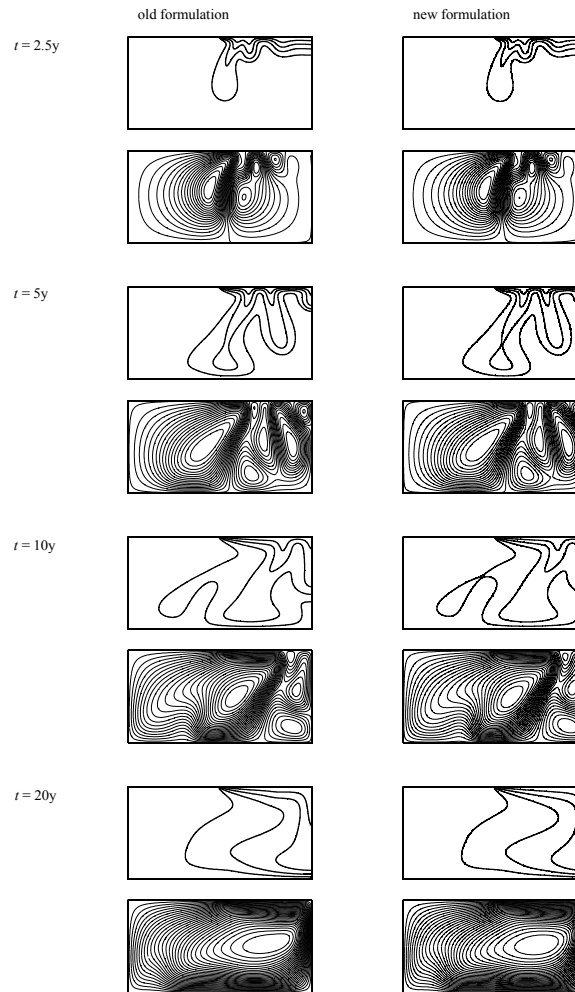
Figure 16.11 Old versus new formulation of the velocity approximation for the Elder problem at grid level $l = 7$: salinities (0.2, 0.4, 0.6 and 0.8 isolines) and streamline patterns for different times t.

16.5.3 The saltpool problem

The saltpool problem has been introduced by Oswald[18]. It represents a three-dimensional saltwater upconing process in a cubic laboratory box under the influence on density and hydrodynamic dispersion. A stable layering of saltwater below freshwater is considered in time for two cases of density: (1) *low* density of 1% mass fraction and (2) *high* density of 10% mass fraction. The problem is defined in Fig. 16.12 and the parameters are listed in Tab. 16.1. A cube of side length 0.2 m is filled with a homogeneous porous medium of porosity ε. At initial time saltwater is layered below freshwater forming a horizontal narrow transition zone. The cubic box is recharged with freshwater through a single inflow hole at a constant rate Q. Through the outflow hole water discharges with a variable salinity. An important outcome of the laboratory experiments are the breakthrough curves of salinity at the outflow hole. It is a challenging task[2] to model these breakthrough behavior.

The problem is difficult caused by very small dispersivities β_L, β_T and a high density contrast particularly for the high density case with a 10% mass fraction of saltwater. Salinity-dependent viscosity effects[4] have to be taken into account too. The mixing concentration at the outflow is measured at a small magnitude amounting in the order to $\frac{1}{100}$ and $\frac{1}{1000}$ related to maximum salinity C_s for the low and the high density case, respectively.
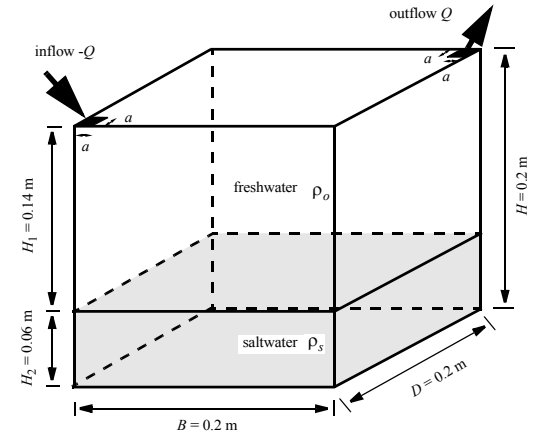
Figure 16.12 Definition of the saltpool problem.

The saltpool problem has been modeled by various authors[2,16,19] with different success. The best agreements with the measurements have been recently achieved by Johannsen *et al.*[16], who used the above new formulation for the consistent velocity approximation in the d^3f code, and additionally, however, adjusted some parameters within accepted bounds. It was shown that very fine meshes (up to about 17 million nodal points) are required to model the high density case with a sufficient accuracy. A hierarchy of regular meshes consisting of hexahedral elements up to grid level $l = 8$ has been studied, where the number of elements is

$$ne = 8^l \tag{16-51}$$

In the present simulation we employ meshes of only moderate sizes as listed in Tab. 16.2. Both a structured

mesh of hexahedral elements with a grid level of 6 and an unstructured mesh of pentahedral elements for only the symmetric half which is partially refined at the outlet are simulated. For the computations the Galerkin-FEM without any upwind and the Adams-Bashforth/ trapezoid rule with adaptive time stepping and one-step Newton are applied, that means the numerical results will be second order accurate both spatially and temporarily.

Table 16.1 Parameters of the saltpool problem

quantity	symbol	magnitude		unit
		low density	high density	
cell height	H	0.2		m
cell width	B	0.2		m
cell depth	D	0.2		m
opening width	a	10^{-3}		m
initial freshwater height	H_1	0.14		m
initial saltwater height	H_2	0.06		m
hydraulic conductivity	$K = \dfrac{k\rho_o g}{\mu_o}$	$97.73 \cdot 10^{-4}$		ms^{-1}
solute expansion coefficient (relative density difference)	$\alpha = \dfrac{\rho_s - \rho_o}{\rho_o}$	$76.0 \cdot 10^{-4}$	$735.0 \cdot 10^{-4}$	1
diffusion coefficient	D_d	$1.0 \cdot 10^{-9}$		$\mathrm{m^2 s^{-1}}$
longitudinal dispersivity	β_L	$1.2 \cdot 10^{-3}$		m
transverse dispersivity	β_T	$1.2 \cdot 10^{-4}$		m
porosity	ε	0.372		1
fluid compressibility	S_o	0.0		m^{-1}
inflow/outflow rate	Q	$1.89 \cdot 10^{-6}$ 0.163296	$1.83 \cdot 10^{-6}$ 0.158112	$\mathrm{m^3 s^{-1}}$ $\mathrm{m^3 d^{-1}}$
variable fluid viscosity[4]	$\mu = \mu_o(1 + 1.85\omega - 4.1\omega^2 + 44.5\omega^3)$ $\omega = C/\rho_o$ mass fraction			

Table 16.2 Used meshes for the saltpool problem
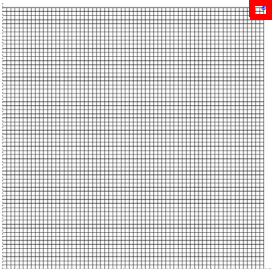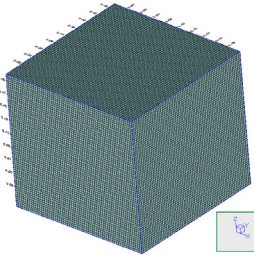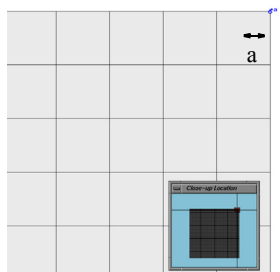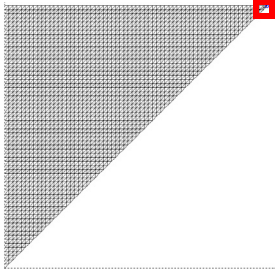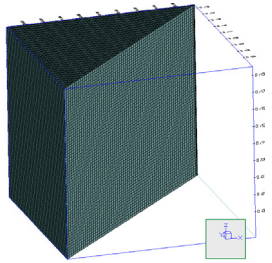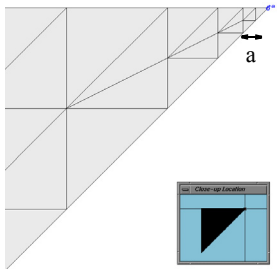
type	view	close-up at the outlet	mesh characteristic
A	regular mesh, hexahedral elements, 2D-view, level $l = 6$ 3D-view		$ne^a = 262{,}144$ $np^b = 274{,}625$ $h_c{}^c = 3.125$ mm $h_d{}^d = 3.125$ mm $h_z{}^e = 3.125$ mm

Table 16.2 Used meshes for the saltpool problem (continued)

type	view	close-up at the outlet	mesh characteristic
B	Irregular mesh, pentahedral elements, 2D-view, symmetric half 3D-view		$ne = 262,912$ $np = 140,010$ $h_c = 3.125$ mm $h_d = 0.552$ mm $h_z = 3.125$ mm

a. total number of elements
b. total number of nodes
c. characteristic horizontal element length in the central region
d. characteristic horizontal element length at outflow/inflow boundaries
e. characteristic vertical element length

The old formulation of the velocity approximation (local smoothing of nonconsistent velocities) completely failed in the saltpool problem for the high density case. This was already observed by Oswald *et al.*[19] where the saltwater mixing concentration at the outlet was significantly overestimated in this case. Such a bad behavior is depicted in Fig. 16.13 showing an overesti-mation of more than 20 times with respect to the experimental salinity for the high density case at the outlet if using the old formulation with the mesh A. On the other hand, the low density case agrees quite well with the experiments if using the old formulation as seen in Fig. 16.13.
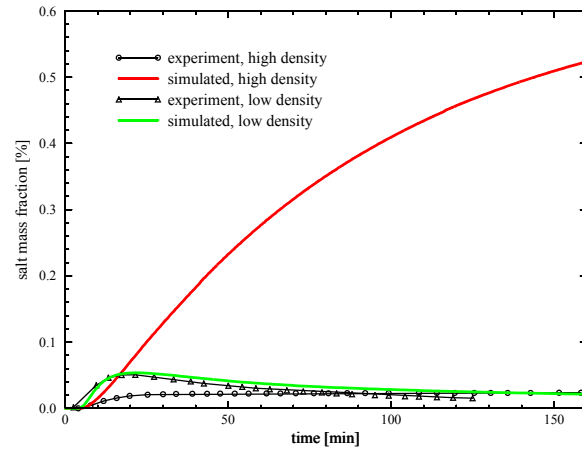
Figure 16.13 Salinity breakthrough curves at the outlet for the high and low density cases: Measured versus simulated salinities for mesh A based on the old formulation of the velocity approximation.

In using the new consistent velocity approximation the computed breakthrough curves are now in reasonable agreement with the experiments as shown in Figs. 16.14 and 16.15 for both the low and the high density case. We note that the parameters are not adjusted during the present simulations. Apparently, an adjustment of the parameters, particularly the transverse dispersivity β_T, porosity ε and conductivity K, is required to attain a better match with the experiments so as performed by Johannsen *et al.*[16]. Otherwise, for the high density case more refined meshes seem to be necessary to improve the breakthrough behavior at the outlet.
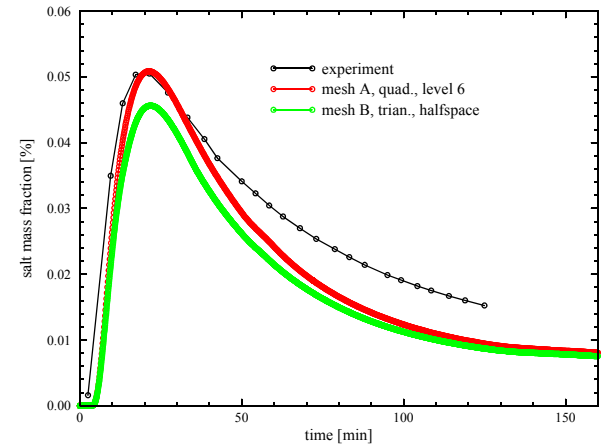
Figure 16.14 Salinity breakthrough curves at the outlet for the low density case: Measured versus simulated salinities for mesh A and B based on the new formulation of the consistent velocity approximation.
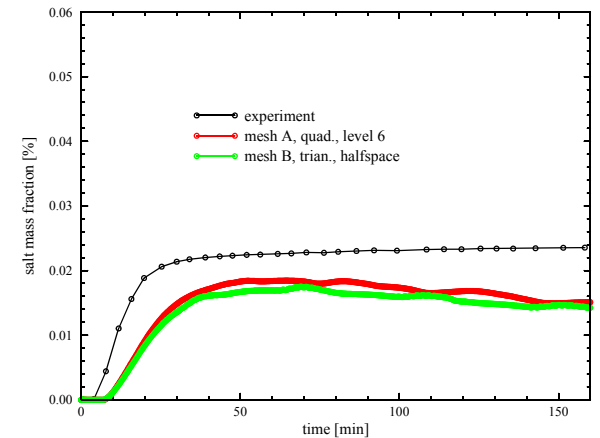
Figure 16.15 Salinity breakthrough curves at the outlet for the high density case: Measured versus simulated salinities for mesh A and B based on the new formulation of the consistent velocity approximation.

Figure 16.16 illustrates the salinity distributions computed by using the new consistent velocity approximation for the low and the high density cases. It reveals the role of the density effects in the mixing and dilution of saltwater controlled mainly by hydrodynamic dispersion. If the case of the high density the transition zone between saline and fresh water is significantly widened forming a 'diffusive upcone' below the outlet, however, at very low concentrations. This mixing process is significantly influenced by the advective and dispersive forces acting locally at the saltwater-freshwater interface which is initially very narrow. A highly accurate and a fully consistent velocity approximation has proven a fundamental requirement for a successful solution of the saltpool problem at high density. Small local inconsistencies in the velocity field would have dramatic consequences on the computational results.

low density   high density
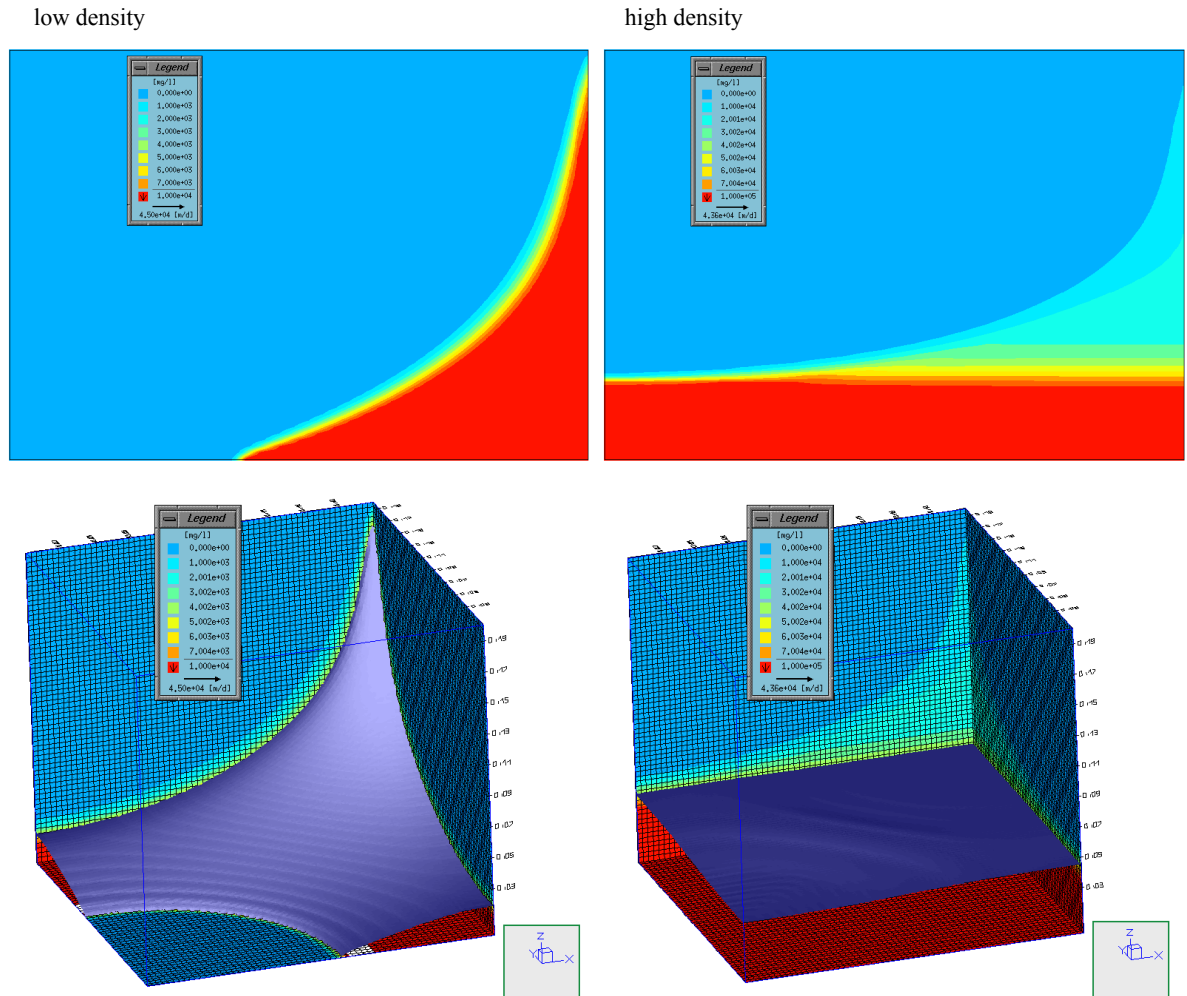
Figure 16.16 Cross-sectional salinity (above) and 50% salinity isosurface (below) at t = 160 min for the low and high density case simulated by the new formulation of the consistent velocity approximation (mesh A).

16.6 Closure

For density-dependent flow and transport processes a new formulation of a consistent velocity approximations developed by Frolkovic[8] and Knabner[12] is incorporated in the FEFLOW code. This became necessary to eliminate errors for high density contrasts arising distinctly in simulating the saltpool problem. The old formulation which is based on a smoothing technique of basically nonconsistent velocities can give lacks in the consistency of velocities at locations of high density gradients. For the saltpool problem at high density it has shown dramatic consequences in the results. Though the saltpool problem is somewhat specific and extreme the new technique is to be recommended (and actually now the default option) for all further density-dependent problems.

The question arises why the lack in the old formulation was not noticed in previous studies (*cf.*,[3-8]). The answer becomes clear in the light of the above comparisons made between the old and the new formulation: Most of the previous studies focussed on moderate density effects and flow situations which are different to that of the saltpool problem where a dispersion-controlled flushing over a narrow transition zone is dominant. Otherwise we should mention that reliable quantitative results for a saltwater mixing process under density effects were not available before Oswald's work[18].

The recomputation of density-dependent problems gives under moderate parameter conditions no remarkable differences between the old and the new formulations so as exemplified above for the Elder problem where the results are in close agreement in both formulations with the recent findings presented by Frolkovic and De Schepper[10]. However, if the density contrasts are much higher and mixing processes over stable narrow saltwater-freshwater transition zones are important only the new formulation of an exact local consistency can guarantee a quantitatively (and physically) correct solution.

References

1. Ackerer, P., Younes, A. & Mose, R., Modeling variable density flow and solute transport in porous medium: 1. Numerical model and verfication, *Transport in Porous Media* **35** (1999) 3, 345-373.

2. Ackerer, P., Younes, A., Oswald, S.E. & Kinzelbach, W., On modelling of density driven flow, *Calibration and Reliability in Groundwater Modelling*, (Proceedings of the ModelCARE 99 Conference held at Zürich, Switzerland, September 20-23, 1999), ed. by F. Stauffer *et al.*, IAHS Publ. no. **265**, 2000, 377-384.

3. Diersch, H.-J., Finite element modelling of recirculating density-driven saltwater intrusion processes in groundwater, *Adv. Water Res.* **11** (1988) 1, 25-43.

4. Diersch, H.-J.G., Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.

5. Diersch, H.-J.G. & 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 held at Golden, Colorado, September 25-27, 1996), ed. by K. Kovar & P. van der Heijde, IAHS Publ. no. 237, 407-415.

6. Diersch, H.-J.G. & Kolditz, O., Coupled groundwater flow and transport: 2. Thermohaline and 3D convection systems, *Adv. Water Res.* **21** (1998), 401-425.

7. Diersch, H.-J.G., Note to the opposing flow regime at mixed convection around a heated cylinder in a porous medium, *Transport in Porous Media* **38** (2000) 3, 345-352.

8. Frolkovic, P., Consistent velocity approximation for density driven flow and transport. in: R. Van Keer *et al.* (eds.), *Advanced Computational Methods in Engineering*, Part 2, Shaker Publishing, Maastrich, 1998, 603-611.

9. Frolkovic, P., *d³f documentation - the discretization.* in: User's manual (compiled by E. Fein), Gesellschaft für Anlagen- und Reaktorsicherheit mbH, Braumschweig, 1998.

10. Frolkovic, P. & De Schepper, H., Numerical modelling of convection dominated transport with density driven flow in porous media, *Adv. Water Res.* **24** (2000) 1, 63-72.

11. Herbert, A.W., Jackson, C.P. & Lever, D.A., Coupled groundwater flow and solute transport with fluid density strongly dependent on concentration, *Water Resour. Res.* **24** (1988), 1781-1795.

12. Knabner, P. & Frolkovic, P., Consistent velocity approximation for finite volume or element discretizations of density driven flow in porous media. in: A.A. Aldama *et al.* (eds.), *Computational Methods in Water Resources XI*, Vol. 1 - *Computational Methods in Subsurface Flow and Transport Problems*, Computational Mechanics Publications, Southampton, 1996, 93-100.

13. Kolditz, O. & Diersch, H.-J., Quasi-steady-state strategy for numerical simulation of geothermal circulation processes in hot dry rock fractures, *Int. J. Non-Linear Mechanics* **28** (1993) 4, 467-481.

14. Kolditz, O., Ratke, R., Diersch, H.-J.G. & Zielke, W., Coupled groundwater flow and transport: 1. Verification of density variable flow and transport models, *Adv. Water Res.* **21** (1998), 27-46.

15. Lee, R.L., Gresho, P.M. & Sani, R.L., Numerical smoothing techniques applied to some finite element solutions of the Navier-Stokes equations. *Proc. 1st Int. Conf. Finite Elements in Water Resources*, (ed. Gray, W. G. *et al.*) Princeton Univ., July 1976, Pentech Press, London, 1977, 4.127-4.145.

16. Johannsen, K., Kinzelbach, W., Oswald, S. & Wittum, G., The saltpool benchmark problem - numerical simulation of saltwater upconing in a porous medium, *Adv. Water Res.* **25** (2002) 3, 335-348.

17. Leijnse, A., *Three-dimensional modeling of coupled flow and transport in porous media.* PhD thesis, University of Notre Dame, Indiana, 1992.

18. Oswald, S.E., Dichteströmungen in porösen Medien: Dreidimensionale Experimente und Modellierung. PhD thesis, ETH Zürich, Switzerland, 1998.

19. Oswald, S.E., Scheidegger, M.B. & Kinzelbach, W., A three-dimensional physical benchmark test for verification of variable-density flow models, submitted to *Water Resour. Res.* June 2000.

20. Voss, C.I., A finite-element simulation model for saturated-unsaturated, fluid-density-dependent groundwater flow with energy transport or chemically reactive single-species solute transport, SUTRA documentation, USGS, Report 84-4369, 1984.

21. Voss, C.I. & Souza, W.R., Variable density flow and solute transport simulation of regional aquifers containing a narrow freshwater-saltwater transition zone. *Water Resour. Res.* **23** (1987), 1851-1866.

22. Yeh, G.T., On the computation of Darcian velocity and mass balance in the finite element modeling of groundwater flow. *Water Resour. Res.* **17** (1981) 1529-1534.

Appendix A

Nomenclature

<u>Latin symbols</u>

C, C_o	ML^{-3}	concentration and reference concentration (salinity), respectively;
C_s	ML^{-3}	maximum concentration;
c	$L^2 T^{-2} \Theta^{-1}$	specific heat capacity of fluid;
$\boldsymbol{D}$	$L^2 T^{-1}$	tensor of mechanical dispersion;
D_d	$L^2 T^{-1}$	molecular diffusion in the porous medium;
E^f, E^s	$L^2 T^{-2}$	internal (thermal) energy density for fluid and solid, respectively;
$\boldsymbol{e}$	1	gravitational unit vector with respect to global coordinates;
$\boldsymbol{e}_{(\xi, \eta, \zeta)}$	1	gravitational unit vector with respect to local coordinates;
f_μ	1	fluid viscosity relation function;

$\boldsymbol{g}$	LT^{-2}	gravity vector in global coordinate directions;
g	LT^{-2}	gravitational acceleration;
H_ξ, H_η, H_ζ	L	integral functions related to local coordinates;
h	L	hydraulic (piezometric) head;
$\boldsymbol{I}$	l	unit vector;
$\boldsymbol{J}$		Jacobian transformation matrix;
$\boldsymbol{J}_C$	$ML^{-2}T^{-1}$	Fickian mass flux vector;
$\boldsymbol{J}_E$	MT^{-3}	Fourierian thermal energy flux vector;
$\boldsymbol{K}$	LT^{-1}	tensor of hydraulic conductivity;
$\boldsymbol{k}$	L^2	tensor of permeability for the porous medium;
l		grid level;
M_{mn}		mass matrix;
N		finite element shape function;
N_m		finite element shape function at node m;
ne		number of elements;
np		number of nodes (points);
n_p		number of nodal contributions at an element patch;
p	$ML^{-1}T^{-2}$	fluid pressure;
Q	L^3T^{-1}	fluid discharge;
Q_ρ	$ML^{-3}T^{-1}$	bulk fluid flow sink/source;
Q_C	$ML^{-3}T^{-1}$	bulk mass sink/source;
Q_E	$ML^{-1}T^{-3}$	bulk thermal sink/source;
S_o	L^{-1}	storage coefficient;
T, T_o	Θ	temperature and reference temperature, respectively;
t	T	time;
$\boldsymbol{v}$	LT^{-1}	Darcy velocity (flux) vector;
x, y, z	L	global Cartesian coordinates;

$\boldsymbol{Z}_m$		RHS vector;
z	L	elevation above a reference datum;

Greek symbols

α	l	solutal expansion coefficient;
β	Θ^{-1}	thermal expansion coefficient;
β_L, β_T	L	longitudinal and transverse dispersivity, respectively;
ε	l	porosity;
θ		variable;
Δ	$MLT^{-3}\Theta^{-1}$	tensor of thermal hydrodynamic dispersion of fluid phase;
λ^f, λ^s	$MLT^{-3}\Theta^{-1}$	thermal conductivity for fluid and solid, respectively;
μ, μ_o	$ML^{-1}T^{-2}$	dynamic viscosity and reference dynamic viscosity of fluid, respectively;
ρ, ρ_o	ML^{-3}	fluid density and reference fluid density, respectively;
$\tilde{\rho}$	l	relative fluid density $(= (\rho - \rho_o)/\rho_o)$;
ρ^s	ML^{-3}	solid density;
ξ, η, ζ	L	local coordinates;
Φ		functional;
Ω		domain;
ω	l	mass fraction;
∂_z	L^{-1}	partial differentiation with respect to the global z-coordinate $\left(= \frac{\partial}{\partial z}\right)$;
∂_ξ	L^{-1}	partial differentiation with respect to the local ξ-coordinate $\left(= \frac{\partial}{\partial \xi}\right)$;
∇	L^{-1}	Nabla (vector) operator with respect to global coordinates;
$\nabla_{(\xi, \eta, \zeta)}$	L^{-1}	Nabla (vector) operator with respect to local coordinates;

Subscripts

m, n	nodal indices;
o	reference value;
p	Gauss-point related;
x, y, z	global coordinate directions;
ξ, η, ζ	local coordinate directions;

Superscripts

e	element;
f	fluid (water) phase;
s	solid phase;

Appendix B

Global smoothing (projection) of discontinuous velocities

A global approximation of the smoothed Darcy velocities can be written as

$$\boldsymbol{v}(x, y, z) = \sum_m N_m \boldsymbol{v}_m \qquad (B1)$$

Assume that we have an unsmoothed (discontinuous) velocity field $\boldsymbol{v}^*(x, y, z)$, where either nonconsistent or consistent velocities occur. Then the smooth function which provides a best fit in the least squares sense over the domain Ω can be obtained from a minimization of the functional

$$\Phi = \int_\Omega \left(\boldsymbol{v} - \boldsymbol{v}^*\right)^2 \Rightarrow \text{Min} \qquad (B2)$$

The minimization procedure

$$\frac{\partial \Phi}{\partial \boldsymbol{v}_m} = \int_\Omega 2(\boldsymbol{v} - \boldsymbol{v}^*)\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{v}_m} = 0 \qquad \text{for} \qquad m = 1, 2, \dots \quad (B3)$$

or

$$\int_\Omega N_m\left(\boldsymbol{v} - \boldsymbol{v}^*\right) = 0 \qquad (B4)$$

results in a system of linear equations to solve the smoothed (continuous) velocities $\boldsymbol{v}$, $viz.$,

$$\sum_n M_{mn} \boldsymbol{v}_n = \boldsymbol{Z}_m \qquad (B5)$$

where M_{mn} represents the mass matrix and $\boldsymbol{Z}_m$ is the RHS involving the unsmoothed relations. They are formed in the finite element assembling procedure as

$$M_{mn} = \int_\Omega N_m N_n \qquad (B6)$$

and, by inserting the Darcy velocity components written for the hydraulic head h from (16-2), as

$$\boldsymbol{Z}_m = \int_\Omega N_m \boldsymbol{v}^* = -\int_\Omega N_m [\boldsymbol{K} f_\mu \cdot (\nabla h + \tilde{\rho}\boldsymbol{e})] \qquad (B7)$$

Note, the least square approximation of global smoothing (B7) is equivalent to a Galerkin weighting procedure[22].

A cost-effective alternative appears if the mass matrix M_{mn} is lumped by an row-sum or diagonal scaling

$$M_{mn} = \begin{cases} \int_{\Omega} N_m (\sum N_n) & m = n \\ & m \neq n \\ 0 \end{cases} \quad \text{(B8)}$$

without need to solve the linear equation system (B5). Mass lumping can be shown to be equivalent to an area-weighted averaging for nodal values.

Local smoothing (projection) of discontinuous velocities

Unlike global smoothing, there is an efficient way to smooth velocity fields by using only individual element information. This is termed as local smoothing[15] and provides a simple nodal averaging based on the number of elements joined at a given node of an element patch. Among several approaches suggested FEFLOW employs following two-step local technique:

(Step 1) The discontinuous velocity in each element e

$$v_p^e = -K^e f_\mu^e \cdot (\nabla h + \tilde{\rho} e) \quad \text{(B9)}$$

is computed at the Gauss points p with given approximations for the head h and density $\tilde{\rho} = \tilde{\rho}(C, T)$ at element level e. Note, the velocity (B9) can also be computed in a consistent approximation as described above (cf. (16-46)).

(Step 2) The values at the Gauss points are assigned to the nearest corner node $p \to m$. Each nodal contribution is summed up and, at the end, the nodal values are averaged by their number of nodal contributions n_p from the patch sharing the node m

$$v_m = \left(\sum_e^{\text{patch}} v_m^e \right) / n_p \quad \text{(B10)}$$

16. Consistent velocity approximation in the finite-element simulation of density-dependent mass and heat transport processes

Coupled groundwater flow and transport: Thermohaline and 3D convection systems

H.-J. G. Diersch[a] & O. Kolditz[b]

[a]*WASY Institute for Water Resources Planning and Systems Research, Berlin, Germany*
[b]*Institute of Fluid Mechanics & Computer Applications in Civil Engineering, University of Hannover, Germany*

ABSTRACT

This work continues the analysis of variable density flow in groundwater systems. It focuses on both thermohaline (double-diffusive) and 3D buoyancy-driven convection processes. The finite-element method is utilized to tackle these complex nonlinear problems in two and three dimensions. The preferred numerical approaches are discussed regarding appropriate basic formulations, balance-consistent discretization techniques for derivative quantities, extension of the Boussinesq approximation, proper constraint conditions, time marching schemes, and computational strategies for solving large systems. Applications are presented for the thermohaline Elder and salt dome problem as well as for the 3D extension of the Elder problem with and without thermohaline effects and a 3D Bénard convection process. The simulations are performed by using the package FEFLOW. Conclusions are drawn with respect to numerical efforts and the appropriateness for practical needs.

Key words: porous media, variable density flow, finite element method, double-diffusive convection, thermohaline convec-tion, three-dimensional Bénard convection

Nomenclature

Latin symbols

A	l	L/d aspect ratio.
B	l	buoyancy ratio (Turner number).
C, C_o	ML^{-3}	concentration and reference concentration, respectively.
C_s	ML^{-3}	maximum concentration.
c^f, c^s	$L^2 T^{-2} \Theta^{-1}$	specific heat capacity of fluid and solid, respectively.
D_d	$L^2 T^{-1}$	medium molecular diffusion coefficient of fluid.
D_{ij}	$L^2 T^{-1}$	tensor of hydrodynamic dispersion.
d	L	thickness (height).
e	L	extent.
e_j	l	components of the gravitational unit vector.
f_μ	l	constitutive viscosity relation function.
g	LT^{-2}	gravitational acceleration.
h	L	hydraulic head.
I	l	e/L symmetric intrusion ratio.
K	LT^{-1}	isotropic hydraulic conductivity constant.
K_{ij}	LT^{-1}	tensor of hydraulic conductivity.
k_{ij}	L^2	tensor of permeability.

L	L	length.
Le	1	Lewis number.
N_m	1	basis (shape) function at node m.
n_i	1	normal unit vector (positive outward).
p^f	$ML^{-1}T^{-2}$	fluid pressure.
Q_C	$ML^{-3}T^{-1}$	sink/source of contaminant mass.
Q_C^R	MT^{-1}	lumped balance flux of solute (positive inward).
Q_{EB}	T^{-1}	extended Boussinesq approximation term.
Q_T	$ML^{-1}T^{-3}$	sink/source of heat.
Q_T^R	ML^2T^{-3}	lumped balance flux of heat (positive inward).
Q_ρ	T^{-1}	sink/source of fluid.
q_C^R	$ML^{-2}T^{-1}$	prescribed normal boundary mass flux (positive outward).
q_i^f	LT^{-1}	Darcy flux of fluid.
$q_{n_T}^{cond}$	MT^{-3}	normal component of the conductive part of the heat flux (positive outward).
$q_{n_C}^{disp}$	$ML^{-2}T^{-1}$	normal component of the dispersive part of the mass flux (positive outward).
$q_{n_C}^{total}$	$ML^{-2}T^{-1}$	normal component of the convective plus dispersive part of the mass flux (positive outward).
q_{n_h}	LT^{-1}	normal component of the Darcy fluid flux (positive outward).
R, R_d	1	specific retardation factor and its time derivative, respectively.
Ra_s, Ra_t	1	solutal and thermal Rayleigh number, respectively.
Ra_c	1	critical Rayleigh number.
S_o	L^{-1}	specific storage coefficient (compressibility).

T, T_o	Θ	temperature and reference temperature, respectively.
V_q^f	LT^{-1}	$\sqrt{q_i^f q_i^f}$ absolute specific Darcy fluid flux.
w, w_m		spatial weighting function and weighting function at node m, respectively.
x_i	L	Cartesian coordinates, Eulerian spatial coordinate vector.

Greek symbols

$\bar{\alpha}$	1	fluid density difference ratio.
α_L, α_T	L	longitudinal and transverse thermodispersivity, respectively.
$\bar{\beta}$	Θ^{-1}	fluid expansion coefficient.
β_L, β_T	L	coefficients of longitudinal and transverse dispersivity of solute, respectively.
Γ		boundary.
γ	1	error tolerance measure.
ΔC	ML^{-3}	concentration difference.
ΔT	Θ	temperature difference.
Δt_n	T	time step width at time plane n.
ε	1	porosity.
ϑ	T^{-1}	chemical decay rate.
Λ	L^2T^{-1}	thermal diffusivity.
λ_{ij}	$MLT^{-3}\Theta^{-1}$	tensor of hydrodynamic thermodispersion.
λ_{ij}^{cond}	$MLT^{-3}\Theta^{-1}$	tensor of thermal conductivity.
λ_{ij}^{disp}	$MLT^{-3}\Theta^{-1}$	tensor of mechanical thermodispersion.
λ^f, λ^s	$MLT^{-3}\Theta^{-1}$	thermal conductivity for fluid and solid, respectively.
μ^f, μ_o^f	$ML^{-1}T^{-2}$	dynamic fluid viscosity and

		reference viscosity, respectively.
ρ^f, ρ^f_o	ML^{-3}	fluid density and reference density, respectively.
ρ^s	ML^{-3}	solid density.
ς	1	$(T-150)/100$ normalized temperature, T in $°C$.
$\chi(C)$		linear (Henry) or nonlinear (Freundlich, Langmuir) sorptivity function.
ω	1	C/ρ^f mass fraction.
Ω		domain.

Subscripts

i, j	spatial Eulerian coordinate (Einstein's summation convention).
l	direction of gravity in the Cartesian coordinate system.
m, n, k	nodal points (Einstein's summation convention).
n	time plane or normal direction.
o	reference value.
p	Gauss point.

Superscripts

e	finite element.
f	fluid phase.
p	predictor value.
R	prescribed boundary value.
s	solid phase.

17.1 Introduction

Thermohaline (or double-diffusive) convection processes are connected with the presence of heterogeneous temperature and concentration fields. Thus, convective currents can arise from heat and salinity gradients acting simultaneously (e.g., Nield[53], Rubin[59], Rubin and Roth[60], Tyvand[73], Trevisan and Bejan[72], Murray and Chen[52], Shen[67], Angirasa and Srinivasan[2], Nield and Bejan[54], Brandt and Fernando[7]). Geophysical applications of thermohaline models can be found for instance in the field of geothermics and waste disposal in salt formations (Evans and Nunn[24]). Thermohaline effects are important for the production of mineralized thermal water, the reinjection of cooled brine into heated deep aquifers connected with geothermal supply technologies, and groundwater movement near salt domes.

Usually, the phenomena of double-diffusive convection (DDC) are related to the presence of both (1) at least, two properties (substances, thermal energy) stratifying the fluid and having different diffusivities and (2) opposing effects on the vertical density gradient[7]. Accordingly, different regimes can be distinguished: A *diffusive regime* occurs if the destabilizing potential comes from the property with the larger diffusivity, e.g., a stable salinity gradient is heated from below. On the other hand, a *finger regime* exists if the driving (destabilizing) forces are caused by the more slowly diffusing property, e.g. hot saline fluid on top of a stable temperature gradient. Both regimes can also appear in a differentiated form referred here to as a *mixed DDC regime* if both properties can destabilize and affect the fluid during the temporal development, e.g., a heavy cool solute sinks down to a region which is

heated from below, so a finger regime at the beginning converges more to a diffusive regime over time.

The first part of the paper by Kolditz *et al.*[43] mainly focused on the verification of numerical schemes against available benchmarks for density-coupled convection processes. Established test examples (e.g., the Henry problem, Elder problem, and salt dome problem) are only 2D and single-diffusive (either mass or heat-driven) convection processes. But even for these academic, seemingly simplistic 2D problems a number of discrepancies appear, still for most recent findings[55]. It has been shown[43] that numerical schemes with their spatial and temporal resolutions can essentially influence computational results. Figure 17.1 recalls the contradictory results for the Elder problem as well as the salt dome test case obtained by different authors. While Elder[23] and the recomputation done by Voss and Souza[75] used obviously overdiffusive schemes on relatively coarse grids, newer findings[43,55] with refined spatial and temporal discretizations reveal convection pattern which are distinctly different from former work. The flow field indicates now a central upwelling rather than downwelling. More dramatically, Oldenburg and Pruess[55] recently presented new results for the salt dome problem (HYDROCOIN level 1 case 5). They believed to achieve much more accurate solutions for this example. But, their results are fully outside of all results known to date (Fig. 17.1b). All the more, their 'swept forward-type' solutions are suspiciously very near to the pure freshwater case without any density coupling, so TOUGH2's results become widely questionable for problems involving velocity-dependent dispersion effects. A possible reason for this discrepancy is recently indicated by the work of Konikow *et al.*[44]. They showed that a salinity pattern of a swept forward type appears if constrained boundary conditions for the salt dome interface are applied (allowing only dispersive release of brine and precluding any convective release of brine). While the study by Konikow *et al.*[44] is more physically motivated it also gives an indication of the importance of a mathematically (numerically) correct handling of boundary conditions for this type of problems, independently of their physical appropriateness or not.

In the past, Galerkin methods, finite differences (FDM) and finite element methods (FEM) have been employed to solve the nonlinear coupled balance equations for variable density groundwater problems in 2D. Pinder and Cooper[57] used the method of characteristics. Finite elements based on a primitive *u-v-p*-variable formulation are utilized by Segol *et al.*[66], Huyakorn and Taylor[38] and Diersch[12,14,15]. However, the subsequent works desisted from primitive variable approaches because their increased accuracy was shown to be in disproportion to the increased numerical effort and inherent restrictions in formulating boundary conditions. Accordingly, standard formulations succeeded which are based on substituting the Darcy law in the primary balance equations. Recent works devoted to this subject are presented, among others, by Frind[26], Diersch *et al.*[16], Voss and Souza[75], Diersch[17], Hassanizadeh and Leijnse[31], Herbert *et al.*[32], Galeati *et al.*[26], Schincariol *et al.*[63], Fan and Kahawita[25], Oldenburg and Pruess[55], Croucher and O'Sullivan[12], Zhang and Schwartz[79], and Kolditz[42]. On the other hand, three-dimensional applications are related to field problems as given by Huyakorn *et al.*[39], Kakinuma *et al.*[40] and Xue *et al.*[76] and do not consider rigorously the density coupling mechanisms. However, there are prior theoretical and numerical works in three-dimensional free

convection problems mostly focused on the (cavity) Horton-Rogers-Lapwood (HRL) problem[54] presented by Holst and Aziz[34], Zebib and Kassoy[78], Straus and Schubert[69,70], Horne[36], Schubert and Straus[64], Caltagi- rone et al.[9], Chan and Banerjee[9], and Beukema and Bruin[6].

a)
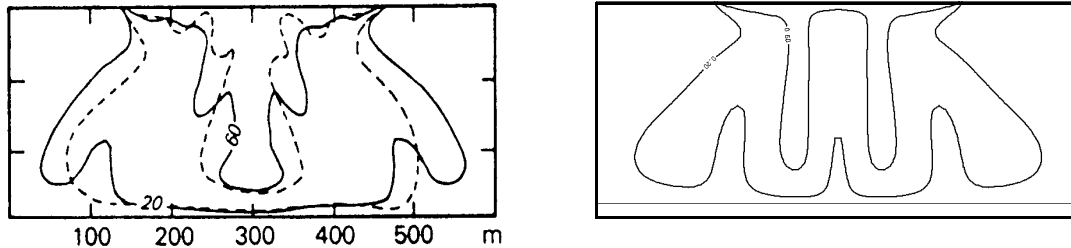
b)
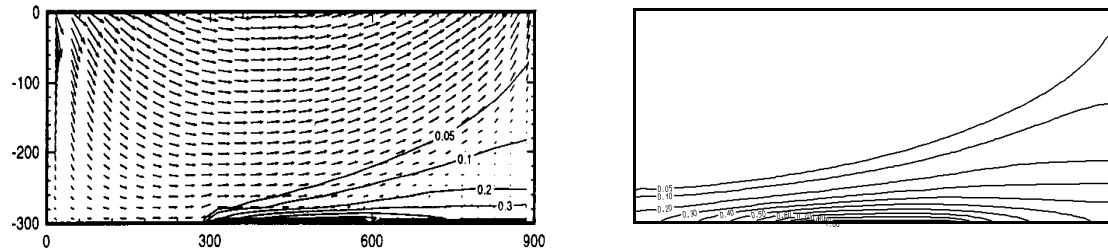
Figure 17.1 a) Simulated concentration pattern at 20 years for the Elder problem with a Rayleigh number of 400: (left) results obtained by the SUTRA simulator (Voss and Souza[75]), (solid curves) and by Elder[23] (dashed curves), and (right) computed by the FEFLOW simulator in agreement with the results attained by ROCKFLOW and TOUGH2 as discussed by Kolditz et al.[43], and b) salt dome test case: (left) TOUGH2 results[55] against (right) FEFLOW (and ROCKFLOW) findings[43] for steady-state with results[55] against (right) FEFLOW (and ROCK-FLOW) findings[43] for steady-state with mechanical dispersion of $\beta_L = 20$ m and $\beta_T = 2$ m.

It is obvious from the above that the extension to thermohaline and/or 3D density-coupled convection problems will significantly increase the importance of both getting a physically equivalent process description in the discretized models and overcoming the numerical burden, particularly if aiming at practical problems.

In the following, relevant numerical aspects are discussed in the context of the FEM. The developed solution strategies are implemented in the 3D finite-element simulator FEFLOW[20]. FEFLOW is employed to study 2D and 3D, thermohaline and buoyancy-driven convection problems from various perspectives.

First, we expand the 2D Elder and salt dome problems to thermohaline processes in order to study thermal influences on groundwater-brine flow systems. Second, we extend the original Elder problem to 3D for both single-diffusive (solutal) and double-diffusive (thermohaline) convection processes to analyze the evolution of 3D pattern formations in comparison with the 2D counterparts. Finally, we devote to a Bénard problem as an example of more complex 3D multicellular convection in a porous layer. The presented results for thermohaline and solutal convection systems may provide examples for a comparison analysis in 2D and 3D by using alternative approaches.

17.2 Basic Equations

The governing equations for the coupled mass and heat transport in groundwater (saturated porous medium) are derived from the basic conservation principles for mass, linear momentum, and energy[43]. The following nonlinear system finally results[20,22] which has to be solved in two and three dimensions

$$S_o \frac{\partial h}{\partial t} + \frac{\partial q_i^f}{\partial x_i} = Q_\rho + Q_{EB}(C, T) \qquad (17\text{-}1)$$

$$q_i^f = -K_{ij} f_\mu \left(\frac{\partial h}{\partial x_j} + \frac{\rho^f - \rho_o^f}{\rho_o^f} e_j \right) \qquad (17\text{-}2)$$

$$\left.\begin{aligned} \frac{\partial}{\partial t}(RC) + \frac{\partial}{\partial x_i}\left(q_i^f C - D_{ij}\frac{\partial C}{\partial x_j} \right) + R\vartheta C &= Q_C \\ &\text{divergent form} \\ R_d \frac{\partial C}{\partial t} + q_i^f \frac{\partial C}{\partial x_i} - \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j} \right) + (R\vartheta + Q_\rho)C &= Q_C \\ &\text{convective form} \end{aligned}\right\} \quad (17\text{-}3)$$

$$[\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s]\frac{\partial T}{\partial t} + \rho^f c^f q_i^f \frac{\partial T}{\partial x_i} - \frac{\partial}{\partial x_i}\left(\lambda_{ij}\frac{\partial T}{\partial x_j} \right) \qquad (17\text{-}4)$$
$$= Q_T$$

To close the set of balance equations the following constitutive formulations are additionally needed:

$$\rho^f = \rho_o^f\left[1 + \frac{\bar{\alpha}}{(C_s - C_o)}(C - C_o) - \bar{\beta}(T - T_o) \right]$$

$$h = \frac{p^f}{\rho_o^f g} + x_l \qquad K_{ij} = \frac{k_{ij}\rho_o^f g}{\mu_o^f} \qquad \bar{\alpha} = [\rho^f(C_s) - \rho_o^f]/\rho_o^f$$

$$f_\mu = \frac{\mu_o^f}{\mu^f(C, T)}$$

$$D_{ij} = (\varepsilon D_d + \beta_T V_q^f)\delta_{ij} + (\beta_L - \beta_T)\frac{q_i^f q_j^f}{V_q^f} \qquad (17\text{-}5)$$

$$R = \varepsilon + (1-\varepsilon)\chi(C) \qquad R_d = \varepsilon + (1-\varepsilon)\frac{\partial[\chi(C)\cdot C]}{\partial C}$$

$$\lambda_{ij} = \lambda_{ij}^{cond} + \lambda_{ij}^{disp} \qquad Q_T = \varepsilon\rho^f Q_T^f + (1-\varepsilon)\rho^s Q_T^s$$

$$\lambda_{ij}^{cond} = [\varepsilon\lambda^f + (1-\varepsilon)\lambda^s]\delta_{ij} \qquad \lambda_{ij}^{disp} = \rho^f c^f\left[\alpha_T V_q^f\delta_{ij} + (\alpha_L - \alpha_T)\frac{q_i^f q_j^f}{V_q^f} \right]$$

As seen a *hydraulic-head-conductivity-(h-K)*-form of the Darcy equation (17-2), instead of the *pressure-permeability-(p-k)*-form, is preferred in FEFLOW which usually permits more convenient formulations of boundary conditions and parameter relations for appli-

cations in subsurface hydrology. As the result, the tensor of hydraulic conductivity K_{ij} refers to the reference density ρ_o^f and the reference viscosity μ_o^f, which are, on the other hand, related to the proper reference conditions for the concentration C_o and the temperature T_o. For such a formulation a *viscosity relation function* f_μ, in eqn (17-5), appears to include viscosity effects in Darcy's law (17-2). The following constitutive polynomial expression is used

$$f_\mu = \frac{\mu_o^f}{\mu^f(C, T)} =$$

$$= \frac{1 + 1.85\omega_{(C = C_o)} - 4.1\omega_{(C = C_o)}^2 + 44.5\omega_{(C = C_o)}^3}{1 + 1.85\omega - 4.1\omega^2 + 44.5\omega^3} \quad (17\text{-}6)$$

$$\times \frac{1 + 0.7063\varsigma - 0.04832\varsigma^3}{1 + 0.7063\varsigma_{(T = T_o)} - 0.04832\varsigma_{(T = T_o)}^3}$$

which is a combination of empirical relationships given by Lever and Jackson[48] for high-concentration saltwater and by Mercer and Pinder[49] for geothermal processes in the range between 0 and 300 °C. In practice, the expansion coefficients $\bar{\alpha}$ and $\bar{\beta}$ of eqn (17-5) are in the most cases considered as constant[54]. For the present investigations we shall also use this assumption to maintain an unified parameter basis for comparison purposes. However, it should be mentioned in a geothermal context where large temperature variations occur and buoyancy forces are dominant, this approach is often not appropriate[56]. Based on the theoretical framework done by Perrochet[56] FEFLOW is also capable of handling a nonlinear variable thermal expansion $\bar{\beta}(T)$ in form of a 5th order polynomial to match the fluid density variation over a wide temperature range

with a high accuracy and to satisfy the zero condition (density anomaly) at 4 °C. For more details see Diersch[22].

The *divergent form* and the *convective form* of the contaminant mass transport equation (17-3) (the energy balance equation (17-4) has already been led to a convective form after introducing the temperature) are physically equivalent. Commonly, the convective form of the transport equation is preferred for numerical approximations because simpler boundary-value problems are accessible.

It is known[28,43] the Boussinesq approximation becomes insufficient for large density variations (e.g., at high-concentration brines or high-temperature gradients). The main difference between the Boussinesq approximation and the actual balance quantities is expressed by the additional term $Q_{EB}(C, T)$ in the continuity equation (17-1) according to

$$Q_{EB}(C, T) = -\underbrace{\varepsilon\left(\frac{\bar{\alpha}}{(C_s - C_o)}\frac{\partial C}{\partial t} - \bar{\beta}\frac{\partial T}{\partial t}\right)}_{1} \quad (17\text{-}7)$$

$$-\underbrace{q_i^f\left(\frac{\bar{\alpha}}{(C_s - C_o)}\frac{\partial C}{\partial x_i} - \bar{\beta}\frac{\partial T}{\partial x_i}\right)}_{2}$$

which is neglected if the Boussinesq approximation is assumed. The first term in eqn (17-7) can be omitted if the temporal changes in concentration and/or temperature vanish. However, even the evolving features of a convection process may be thoroughly affected at higher density contrasts (problems of bifurcation, phys-

ical instability and hydrodynamic pattern formation). The second term of eqn (17-7) can be ignored if the density gradient is essentially orthogonal to the velocity vector. This is quite often not a tolerable assumption. Note, the expression (17-7) has to be modified in the case a nonlinear variable thermal expansion $\bar{\beta}(T)$ [22].

17.3 Spatial Discretization

The above equations (17-1) to (17-4) are discretized by the FEM using bilinear or biquadratic elements for 2D, and prismatic pentahedral trilinear or hexahedral trilinear and triquadratic elements for 3D. Finally, it yields the following coupled matrix system:

$$
\begin{aligned}
O\dot{h} + S(h, C, T)h &= F(h, q, C, \dot{C}, T, \dot{T}) \\
Aq &= B(h, C, T) \\
P(C)\dot{C} + D(q, C)C &= R(C) \\
U\dot{T} + L(q, T)T &= W(T)
\end{aligned}
\tag{17-8}
$$

where h, q, C and T represent the resulting vectors of nodal hydraulic head, Darcy fluxes, contaminant concentration and temperature, respectively. The superposed dot means differentiation with respect to time t. The matrices S, A, O, P and U are symmetric and sparse, while D and L are unsymmetric and sparse. The remaining vectors F, B, R and W encompass the right-hand sides (RHS) of eqns (17-1) to (17-4), respectively. The main functional dependence is shown in parenthesis.

The individual finite-element formulations of the matrix system (17-8) as realized in FEFLOW are summarized in Appendix A. Note, different formulations result for the divergent and the convective forms of the transport equations. Though physically equivalent, they can deliver different numerical solutions due to their different boundary-value formulations.

Another point of view is related to the numerical evaluation of the Darcy fluxes q for a given discretization. The success of a numerical solution for variable density flow problems is essentially dependent on an appropriate choice of suitable schemes for computing derivative quantities from the Darcy equation.

17.4 Continuous Approximation of Velocity Fields

The substitution of Darcy fluxes (17-2) in the continuity equation (17-1) gives immediately an equation to determine the unknown hydraulic head h according to the weak formulation (A3) in Appendix A. If h is known and assuming initial C and T distributions, the fluxes q can be directly computed via Darcy's equation (17-2). However, a careful handling of derivative quantities is required. As normally done in FEM, piecewise continuous ($C°$) basis functions N_m (Appendix A) for the hydraulic head h generate velocity fields q (using derivatives of hydraulic head) that exhibit discontinuities across element boundaries. It results in nonunique values at nodal points. Particularly for buoyancy-influenced flows, discontinuous (nonunique) velocities can cause difficulties (spurious vertical velocities) in the numerical solution due to inappropriate balance approximation of the lower order term $\partial h / \partial x_j$, behaving constantly in an element for the case of linear basis functions, and the higher order gravitational term $[\rho^f(C, T) - \rho_o^f] / \rho_o^f$, varying linearly in an element for

linear basis functions, of the RHS of Darcy equation (17-2). This has already been addressed in previous works[29,32,75] and different numerical schemes were proposed to overcome these problems.

Voss and Souza[75] preferred for the SUTRA code in 2D a reduced order approximation of the buoyancy term, actually the concentration is averaged in every element, therefore, the pressure gradient and the concentration distribution have the same spatial variability, practically constant (for linear basis functions). This is called a consistent velocity evaluation. Leijnse[46] showed that such a consistent velocity approximation can be interpreted as an average of the local gravity component in the local directions of a finite element. A generalization of this spatial averaging has been recently presented by Knabner and Frolkovic[41]. Instead of reducing the approximations Herbert *et al.*[32] introduced a mixed interpolation strategy in NAMMU for 2D, where the pressure is approximated by quadratic elements to obtain a linearly distributed pressure gradient which becomes consistent with a linear distribution of the concentration-dependent buoyancy term. Clearly, quadratic basis functions increase the computational expense and, especially for 3D, an alternative approach is preferable.

Taking into consideration that the discretized balance terms of the conservation equations provide generally a different spatial variability (compare the 'diffusion' term against the 'convective' term or a 'reactive' term), a consistent approximation by the FEM means that all terms have to be rigorously weighted at nodal points. As the result, unique values of even discontinuous variables are generated at nodal points. This principle is consequently applied also to

the velocity evaluation and leads to approaches referred to as *smoothing techniques* used in FEFLOW for the present analysis. Lee *et al.*[45] thoroughly discussed both global and local smoothing techniques for derivative quantities. In this light, the weak form of the Darcy equation (A4) in Appendix A can be recognized as a global smoothing procedure which was introduced in the water resources literature by Yeh[77]. Today, smoothing techniques have an additional meaning for adaptive methods to compute higher order solutions for an error estimation[21]. Appendix B summarizes the smoothing techniques available in FEFLOW and appropriate for the present simulations of coupled phenomena. While global derivative smoothing schemes with a consistent mass matrix require a higher numerical effort, lumped mass smoothing algorithms as well as simpler local smoothing schemes are the most cost-effective approaches and have shown to be well-suited for the present class of problems. The latter is to be recommended for large 3D problems.

Smoothed velocities of a higher-order approximation lead to a continuous distribution of all velocity components in a mesh. As a consequence, continuous fields also exist along material interfaces, e.g., between media with different hydraulic conductivities, where an interfacial nodal point shares these different media and, naturally, a weighted average of the flux quantities results. Leijnse[46] pointed out that physically unrealistic results can be obtained for cases where the conductivity in adjacent elements differ by more than two orders of magnitude. Indeed, if utilizing such continuous velocity fields from a mesh having an insufficiently adapted interface discretization particle tracking procedures can lead to poor results if starting pathlines near such an interface location (a particle may effectively be

propagated into media with low hydraulic conductivity). On the other hand, a discontinuous velocity field approximation often gives significant problems when a particle crosses an element. Then, particle can be 'caught' in the interface due to components which have opposite directions across an element edge as indicated by Sauter and Beusen[62], who introduced special transition elements with interpolated (smoothed) velocity properties to overcome these difficulties. As the sum, the higher-order approximation of continuous velocities is the most natural approach in the finite element method and need not any *ad-hoc* techniques in adapting interface conditions, provided, however, the interface is appropriately discretized. The necessity for a continuous flow field approximation also in the context of modeling heterogeneous media is thoroughly discussed in the work about mixed hybrid finite element techniques presented by Mosé *et al.*[51] followed by recent discussions given by Cordes and Kinzelbach[11] and Ackerer *et al.*.

17.5 Constraints and Related Budget Analysis

Constraints of boundary conditions can play an important role in practical modeling of variable density transport. Typically in saltwater enroachment problems, the boundary conditions of freshwater and saltwater are dependent on the in/outflowing characteristics essential to a correct mathematical formulation. However, most prior works[26,32,38,44,55,66,75] did not consider such conditions in a rigorous manner. To identify the problem let us consider, for instance, the salt dome flow problem as schematized in Fig. 17.2.
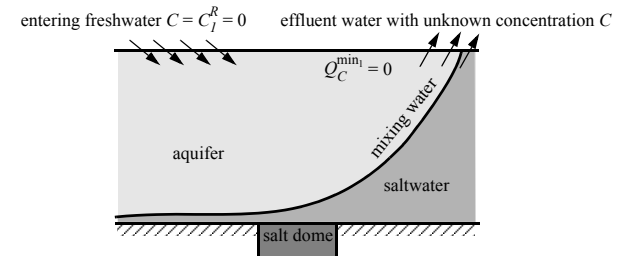
Figure 17.2 Application of transport constraints for saltwater intrusion in flowing groundwater over a salt dome.

Alternating boundary concentrations appear on the top boundary depending on the dynamic process. As long as water enters the domain it should have a prescribed concentration of freshwater. However, if the water leaves the domain (along the same upper boundary) the concentration on this boundary is unknown and should be computed. Such a description can be easily realized if the entire boundary section is assigned by a freshwater boundary condition of 1st kind ($C = C_1^R$), and at the same time, the boundary will be imposed by a constraint condition in form of a null minimum mass flux $Q_C^{min_1} \equiv 0$. Such an arrangement guarantees that the freshwater condition remains valid as long as the convective mass flux, being concentration-dependent due to the density variation, points into the domain.

A rigorous handling of such constraints is permitted by a prescription of *complementary* conditions for each boundary type[20,22]. For instance, the minimum and maximum constraints of a Dirichlet-type concentration will lead to additional conditions in the following form (it reads: the imposed boundary condition $C = C_1^R(t)$ is

accepted only if the related mass balance flux Q_C^R (and the related hydraulic head h^R) is within given min-max bounds, if not, these bounds have to be used as new boundary conditions, where the boundary type has to be changed from a 1st kind into a flux-type boundary condition of a point sink/source Q_C)

$$
\text{1st kind } C_1^R(t) \text{ only if } \left\{ \begin{array}{c} Q_C^R < Q_C^{max_1}(t) \\ \text{and} \\ Q_C^R > Q_C^{min_1}(t) \\ \text{and} \\ h^{min_1} \leq h^R \leq h^{max_1} \end{array} \right\}
$$

else set Q_C as an intermediate flux-type condition according to: $\Bigg\}$ (17-9)

$$
Q_C = \left\{ \begin{array}{ll} Q_C^{max_1}(t) & \text{if } \left\{ Q_C^R \geq Q_C^{max_1}(t) \text{ and } h^{min_1} \leq h^R \leq h^{max_1} \right\} \\ \\ Q_C^{min_1}(t) & \text{if } \left\{ Q_C^R \leq Q_C^{min_1}(t) \text{ and } h^{min_1} \leq h^R \leq h^{max_1} \right\} \\ \\ 0 & \text{if } \{ h^R < h^{min_1} \text{ or } h^R > h^{max_1} \} \end{array} \right.
$$

where Q_C^R is the mass balance flux at the boundary point to be computed while the C_1^R condition is imposed, $Q_C^{max_1}$ and $Q_C^{min_1}$ denote the prescribed time-dependent maximum and minimum bounds, respectively, and Q_C represents a singular mass sink/source to be set at the boundary point (node) instead of the original 1st kind boundary condition. Similar expressions exist for the other types of boundary conditions. This procedure allows the control of concentration at the boundary in dependence on both the balanced flow conditions through the boundary (e.g., $Q_C^{min_1} \equiv 0$) and the location of possible free-surface conditions within the bounds h^{min}, h^{max}. The latter is very important for complex mine flooding processes as studied by Diersch et al.[19].

The computed fluxes Q_C^R represent *lumped* (summed-up) mass balance fluxes at nodal points

$$
Q_C^R = -\int_\Gamma q_C^R \tag{17-10}
$$

Note, the balance quantities are defined positive inward on Γ. Actually, the specific balance fluxes q_C^R are composed by their convective and dispersive parts according to

$$
q_C^R = \underbrace{C^R q_{n_h}^R}_{\text{convective}} - \underbrace{D_{ij}\frac{\partial C}{\partial x_j}n_i}_{\text{dispersive}} \tag{17-11}
$$

In practice, it has been shown to be inappropriate to include the total (convective plus dispersive) flux into the procedure of controlling the constraint conditions because the direction of dispersive fluxes is ambiguous (e.g., the dispersive spreading also occurs against the flow direction). Accordingly, the balance-based evaluation of fluxes is exclusively related to the convective mass fluxes:

$$
Q_C^R = -\int_\Gamma q_C^R \approx -\int_\Gamma (C^R q_{n_h}^R) \tag{17-12}
$$

giving unambiguously directional balance quantities. Similar expressions can be obtained for the balance of convective heat flux, viz.,

$$
Q_T^R \approx -\int_\Gamma \left(\rho^f c^f T^R q_{n_h}^R \right) \tag{17-13}
$$

The computation of the convective part of balance fluxes at each controlling (nodal) point is performed via a budget analysis in a postprocessing step. The basic formulation used for computing the above balance quantities is derived in Appendix C.

17.6 Temporal Discretization and Iterative Solution Process

In general, for more complex flow processes it cannot be predicted which time steps are allowable with respect to the accuracy requirements. Accordingly, a predefined time step marching strategy is often inappropriate and inefficient. Alternatively, stable fully implicit and semi-implicit two-step techniques known as the GLS-(Gresho-Lee-Sani) predictor-corrector time integrator[6,30] with automatically controlled time stepping of first order by the *Forward Euler/Backward Euler* (FE/BE) and of second order by the *Adams-Bashforth/Trapezoid Rule* (AB/TR) have proven to be powerful and accurate strategies, especially for strong nonlinearities and complex situations. At each time step, the convergence tolerance γ directly governs the time-step size. It provides a cost-effective method in that the step size is increased whenever possible and decreased only when necessary due to the error estimates. The GLS scheme is thoroughly described elsewhere[6,17,18,30]. Here, we will only address modified features which are important in the context of the multiple coupling of equations and constraint computation for the present tasks. Note, a full Newton method is embedded into the AB/TR and FE/BE predictor-corrector methods. The overall adaptive solution process is

outlined in Fig. 17.3.
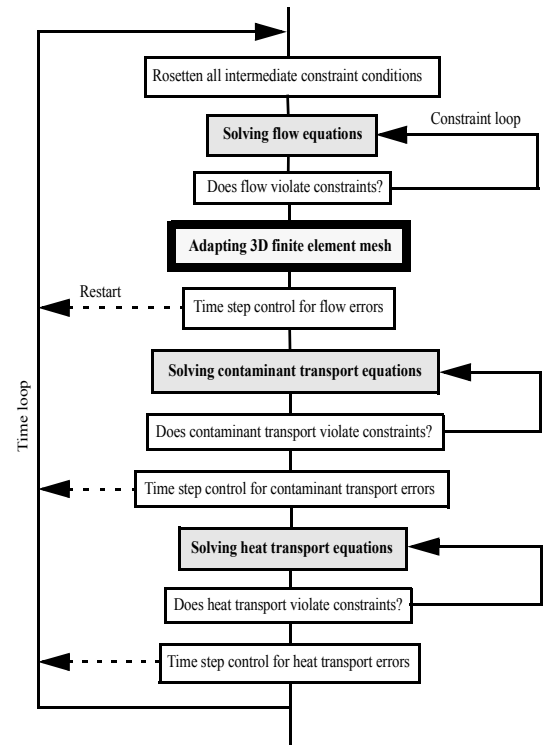
Figure 17.3 Adaptive strategy for coupled transient flow, mass and heat transport.

Denoting the time plane by the subscript n and the variable time step width by Δt_n the coupled matrix system (17-8) is solved in the following 22 raw working steps:

(Step 0) Compute the initial acceleration vectors $\dot{\boldsymbol{h}}_n$, $\dot{\boldsymbol{C}}_n$ and $\dot{\boldsymbol{T}}_n$ for $n = 0$ (once per problem)

$$O\dot{h}_n = F(h_n, C_n, T_n) - S(h_n, C_n, T_n)h_n$$

$$P(C_n)\dot{C}_n = R(C_n) - D(h_n, C_n, T_n)C_n \qquad (17-14)$$

$$U\dot{T}_n = W(T_n) - L(h_n, C_n, T_n)T_n$$

and guess an initial time step Δt_0.

(Step 1) Perform explicit predictor solutions by using the AB and FE algorithm, respectively:

$$\left. \begin{aligned} h^p_{n+1} &= f_{AB}(\Delta t_n, \Delta t_{n-1}, \dot{h}_n, \dot{h}_{n-1}) \\ C^p_{n+1} &= f_{AB}(\Delta t_n, \Delta t_{n-1}, \dot{C}_n, \dot{C}_{n-1}) \\ T^p_{n+1} &= f_{AB}(\Delta t_n, \Delta t_{n-1}, \dot{T}_n, \dot{T}_{n-1}) \end{aligned} \right\} \qquad (17-15a)$$

$$\left. \begin{aligned} h^p_{n+1} &= f_{BE}(\Delta t_n, \dot{h}_n) \\ C^p_{n+1} &= f_{BE}(\Delta t_n, \dot{C}_n) \\ T^p_{n+1} &= f_{BE}(\Delta t_n, \dot{T}_n) \end{aligned} \right\} \qquad (17-15b)$$

The detailed description of the functions $f_{AB}(\ )$ and $f_{BE}(\ )$ can be found in Gresho *et al.*[30], Bixler[6], and Diersch[17,18].

(Step 2) Do corrector solution for the flow equation achieved by the TR and BE scheme, respectively:

$$\left(\frac{2O}{\Delta t_n} + S(h^p_{n+1}, C^p_{n+1}, T^p_{n+1})\right)h_{n+1} \qquad (17-16a)$$

$$= O\left(\frac{2}{\Delta t_n}h_n + \dot{h}_n\right) + F(h^p_{n+1}, C^p_{n+1}, T^p_{n+1}, q_n, \dot{C}_n, \dot{T}_n)$$

$$\left(\frac{O}{\Delta t_n} + S(h^p_{n+1}, C^p_{n+1}, T^p_{n+1})\right)h_{n+1} \qquad (17-16b)$$

$$= \frac{O}{\Delta t_n}h_n + F(h^p_{n+1}, C^p_{n+1}, T^p_{n+1}, q_n, \dot{C}_n, \dot{T}_n)$$

(Step 3) If constraint conditions are violated update the matrix system (17-16a), (17-16b) for the new flow boundary values and restart the flow solution with step 2. If all constraint limits are satisfied continue with step 4.

(Step 4) Solve Darcy equation:

$$Aq_{n+1} = B(h_{n+1}, C^p_{n+1}, T^p_{n+1}) \qquad (17-17)$$

(Step 5) Update the new accelerations vectors by 'inverting' the TR and BE, respectively:

$$\left. \begin{aligned} \dot{h}_{n+1} &= \frac{2}{\Delta t_n}(h_{n+1} - h_n) - \dot{h}_n \\ \dot{h}_{n+1} &= \frac{1}{\Delta t_n}(h_{n+1} - h_n) \end{aligned} \right\} \qquad (17-18)$$

(Step 6) Compute the local truncation error of the approximate flow equation for the AB/TR and FE/BE scheme, respectively:

$$\left. \begin{aligned} d^{flow}_{n+1} &= \frac{h_{n+1} - h^p_{n+1}}{3\left(1 + \frac{\Delta t_{n-1}}{\Delta t_n}\right)} \\ d^{flow}_{n+1} &= \frac{1}{2}(h_{n+1} - h^p_{n+1}) \end{aligned} \right\} \qquad (17-19)$$

(Step 7) Predict the potential new time step length from the error estimates of the flow equation:

$$\Delta t_{n+1}^{flow} = \Delta t_n \left(\frac{\gamma}{\left\| d_{n+1}^{flow} \right\|} \right)^{1/\kappa} \qquad (17\text{-}20)$$

where κ is 3 for the AB/TR and 2 for the FE/BE scheme, γ is a user-specified error tolerance ($\gamma = 10^{-4} - 10^{-3}$ is typical), and $\| \cdot \|$ is a norm to be chosen as the weighted RMS

$$\left\| d_{n+1}^{flow} \right\| = \left[\frac{1}{NP} \left(\frac{1}{h_{max}} \sum_i (h_{i(n+1)} - h_{i(n)})^2 \right) \right]^{1/2} \qquad (17\text{-}21)$$

or, alternatively, as the maximum norm

$$\left\| d_{n+1}^{flow} \right\| = \frac{\max_i |h_{i(n+1)} - h_{i(n)}|}{h_{max}} \qquad (17\text{-}22)$$

in which NP is the total number of points and h_{max} corresponds to the maximum value of the hydraulic head.

(Step 8) Tactics for acceptance of the predicted new time step: If the flow solution does not satisfy the prescribed accuracy the time step is reduced by using appropriate formulae[17,18] and the flow solution is restarted with step 2. Otherwise, if the accuracy is satisfied the solution process is continued with step 9.

(Step 9) Perform corrector solution for the mass transport equation achieved by the TR and BE scheme, respectively:

$$\left. \begin{array}{l} \left(\dfrac{2P(C_{n+1}^p)}{\Delta t_n} + D(q_{n+1}, C_{n+1}^p) + J_p(C_{n+1}^p) \right) C_{n+1} = P(C_{n+1}^p) \left(\dfrac{2}{\Delta t_n} C_n + \dot{C}_n \right) + J_p(C_{n+1}^p) C_{n+1}^p + R(C_{n+1}^p) \\[2em] \left(\dfrac{P(C_{n+1}^p)}{\Delta t_n} + D(q_{n+1}, C_{n+1}^p) + J_p(C_{n+1}^p) \right) C_{n+1} = \dfrac{P(C_{n+1}^p)}{\Delta t_n} C_n + J_p(C_{n+1}^p) C_{n+1}^p + R(C_{n+1}^p) \end{array} \right\} \qquad (17\text{-}23)$$

where $J_p(C_{n+1}^p)$ is the partial (tangential) Jacobian matrix based on the predictor which results from the embodied full Newton approach. Its specific expressions depend on the divergent and convective form of the used transport equation as given by Diersch[17].

(Step 10) If mass constraint conditions are violated update the matrix system (17-23) for the new mass boundary values and restart the mass solution with step 9. Otherwise, continue with step 11.

(Step 11) Update the new acceleration vectors $\dot{C}_{n+1}$ for the concentration similar to step 5.

(Step 12) Equivalently to step 6 compute the local truncation error of mass transport d_{n+1}^{mass} based on $(C_{n+1} - C_{n+1}^p)$.

(Step 13) Estimate the potential new time step from the mass transport computation Δt_{n+1}^{mass}, similar to step 7 by using the error d_{n+1}^{mass}.

(Step 14) Accuracy check of mass transport: reject the current mass transport solution and restart at step 2 with a reduced time width Δt_n if the required accuracy could not be satisfied. Otherwise, continue with the heat transport solution at step 15.

(Step 15) Perform corrector solution for the heat transport equation accomplished by the TR and BE scheme, respectively:

$$\left.\begin{array}{l} \left(\dfrac{2U}{\Delta t_n} + L(q_{n+1}, T_{n+1}^p) + J_p(T_{n+1}^p)\right)T_{n+1} = U\left(\dfrac{2}{\Delta t_n}T_n + \dot{T}_n\right) + J_p(T_{n+1}^p)T_{n+1}^p + W(T_{n+1}^p) \\[3mm] \left(\dfrac{U}{\Delta t_n} + L(q_{n+1}, T_{n+1}^p) + J_p(T_{n+1}^p)\right)T_{n+1} = \dfrac{U}{\Delta t_n}T_n + J_p(T_{n+1}^p)T_{n+1}^p + W(T_{n+1}^p) \end{array}\right\} \qquad (17\text{-}24)$$

(Step 16) If heat constraint conditions are violated update the matrix system (17-24) for the new heat boundary values and restart the heat transport solution with step 15. Otherwise, continue with step 17.

$$\Delta t_{n+1} = min(\Delta t_{n+1}^{flow}, \Delta t_{n+1}^{mass}, \Delta t_{n+1}^{heat}) \qquad (17\text{-}25)$$

and restart the time loop with step 1 as long as the final time is not reached.

(Step 17) Update the new accelerations vectors $\dot{T}_{n+1}$ for the temperature similar to step 5.

(Step 18) Compute the local truncation error of heat transport d_{n+1}^{heat} based on $(T_{n+1} - T_{n+1}^p)$.

(Step 19) Estimate the potential new time step from the heat transport computation Δt_{n+1}^{heat}, similar to step 7 by employing the error d_{n+1}^{heat}.

(Step 20) Accuracy check of heat transport: reject the current heat transport solution and restart with step 2 for a reduced time step if the required accuracy could not be satisfied. Otherwise, continue with step 21.

(Step 21) Determine the new time step length

As seen above a constraint violation can lead to recycling steps around the matrix solution process for flow, mass and heat transport. The matrix updating gains efficiency if a total reassembly can be avoided. Such a procedure of constraint feedback is generally not restricted in the number of loops. Normally, if constraint conditions are raised two recycles become sufficient.

To solve the resulting large sparse matrix systems ((17-14), (17-16a), (17-16b), (17-17), (17-23), (17-24)) appropriate iterative solvers for symmetric and unsymmetric equations have to be applied[3]. For the symmetric positive definite flow equations the conjugate gradient (CG) method[33] is successful provided a useful

preconditioning is applied. Standard preconditioner such as the incomplete factorization (IF) technique[49] and alternatively a modified incomplete factorization (MIF) technique[4] based on the Gustafsson algorithm are used. Different alternatives are available for the CG-like solution of the unsymmetric transport equations: a restarted ORTHOMIN[5] (orthogonalization-minimization) method, a restarted GMRES[61] (generalized minimal residual) technique and Lanczos-type methods[47,71], such as CGS[68] (conjugate gradient square), BiCGSTAB[74] (bi-conjugate gradient stable) and BiCGSTABP[74] (postconditioned bi-conjugate gradient stable). For preconditioning an incomplete Crout decomposition scheme is currently applied. Commonly, BiCGSTABP is the first choice in our practical simulation of large problems.

17.7 Examples of 2D Thermohaline Systems

17.7.1 Dimensionless parameters

From a dimensional analysis of the governing balance equations one can derive the following dimensionless parameters[54] to characterize the convection processes:

solutal Rayleigh number Ra_s :

$$Ra_s = \frac{\frac{\overline{\alpha}}{(C_s - C_o)} \cdot \Delta C \cdot K \cdot d}{\varepsilon \cdot D_d} \qquad (17\text{-}26)$$

thermal Rayleigh number Ra_t :

$$Ra_t = \frac{\overline{\beta} \cdot \Delta T \cdot K \cdot d}{\Lambda} \qquad \Lambda = \frac{\varepsilon \lambda^f + (1-\varepsilon)\lambda^s}{\rho^f c^f} \qquad (17\text{-}27)$$

Lewis number Le :

$$Le = \frac{\Lambda}{\varepsilon \cdot D_d} \qquad (17\text{-}28)$$

Buoyancy ratio (Turner number) B :

$$B = \frac{\frac{\overline{\alpha}}{(C_s - C_o)} \cdot \Delta C}{\overline{\beta} \cdot \Delta T} \qquad (17\text{-}29)$$

Accordingly, the relation between the solutal and thermal Rayleigh number is given by

$$Ra_s = B \cdot Le \cdot Ra_t \qquad (17\text{-}30)$$

From perturbation analysis along the thermohaline Horton-Rogers-Lapwood (HRL) problem[54] the critical Rayleigh number Ra_c is composed of solutal and thermal influences. It can be shown for the HRL problem that boundary between stable and instable convection possesses a straight line, viz.,

$$Ra_c = Ra_s + Ra_t \qquad (17\text{-}31)$$

The critical Rayleigh number Ra_c depends on boundary conditions, geometry and anisotropy. A first critical number Ra_{c1} describes the onset of convection in the form of stable stationary rolls which is normally given by $4\pi^2$. Further increase of the Rayleigh number leads to a second critical stage characterized by Ra_{c2}. For

this regime no more stationary conditions exist and fluctuating (oscillatory) transient convective patterns appear. Ra_{c2} is only known from numerical studies[35,37,58,65], where a value of about 390 is reported. For 3D cases it has been found the final convective structures are dependent on the initial conditions. Stable convection could be recognized only if raised as 2D roll cells. Otherwise, the 3D state has found to be instable from the beginning[78] as soon above criticality.

17.7.2 The 2D thermohaline Elder problem

17.7.2.1 Definition of the problem

The 2D saline Elder problem[43] is expanded to a thermohaline convection process if the salinity field is augmented by a thermal distribution as defined in Fig. 17.4. The geometry is given by the aspect ratio $A = L/d$ of 4 and a so-called intrusion ratio $I = e/L$ of 0.5. While the homogeneous aquifer is permanently heated from below, the salinity gradient acts from above. The normalized concentration on the top of the aquifer is greater than zero in the central section. On the bottom of the aquifer the salinity is held at zero. On the other hand, the top and bottom boundaries are held at constant temperatures as indicated in Fig. 17.4. Otherwise, all remaining boundary portions are considered impervious for solute and adiabatic (insulated) for heat. All boundaries are impervious for fluid flow. As a reference for the hydraulic head a single boundary value of $h = 0$ has to be set at one node (normally in the centre of the mesh). The used model parameters are summarized in Tab.17.1.

As stated above, such a formulation of the thermohaline Elder problem can be considered as a mixed DDC regime where a finger regime dominates at the beginning (cool salinity sinks down) and later a more diffusive regime occurs (downsunk salinity is heated from below).
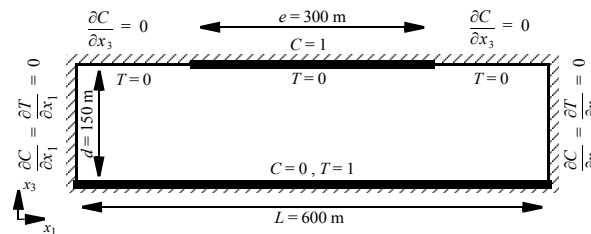
Figure 17.4 Definition of the 2D thermohaline Elder problem (modified from Voss and Souza[75]).

Table 17.1 Simulation parameters for the 2D thermohaline Elder problem

Symbol	Quantity	Value	Unit
A	aspect ratio	4.	1
B	buoyancy ratio (Turner number)	1, 2, 3, 4, 5	1
C_0	reference concentration	0.	g l^{-1}
$c^f \rho^f$	thermal capacity of fluid	$4.2 \cdot 10^6$	J m^{-3} K^{-1}

Table 17.1 Simulation parameters for the 2D thermohaline Elder problem (continued)

Symbol	Quantity	Value	Unit
D_d	molecular diffusion coefficient	$3.565 \cdot 10^{-6}$	$m^2\ s^{-1}$
d	thickness (height)	150	m
e	extent of intrusion	300	m
f_μ	viscosity relation function	1	1
I	symmetric intrusion ratio	0.5	1
K	hydraulic conductivity	$4.753 \cdot 10^{-6}$	$m\ s^{-1}$
L	length	600	m
Le	Lewis number	1	1
Ra_s	solutal Rayleigh number	400	1
Ra_t	thermal Rayleigh number	400, 200, 133.3, 100, 80	1
T_0	reference temperature	0.	K
ΔT	temperature difference	400, 200, 133.3, 100, 80	K
α_L	longitudinal thermodispersivity	0.	m
α_T	transverse thermodispersivity	0.	m
$\bar{\alpha}\,/\,C_s$	density ratio	0.2	1

Table 17.1 Simulation parameters for the 2D thermohaline Elder problem (continued)

Symbol	Quantity	Value	Unit
β_L	longitudinal dispersivity of solute	0.	m
β_T	transverse dispersivity of solute	0.	m
$\bar{\beta}$	thermal expansion coefficient	$5 \cdot 10^{-4}$	K^{-1}
ε	porosity	0.1	1
Λ	thermal diffusivity	$3.565 \cdot 10^{-7}$	$m^2\ s^{-1}$
λ^f	thermal conductivity of fluid	0.65	$J\ m^{-1}\ s^{-1}\ K^{-1}$
λ^s	thermal conductivity of solid	1.591	$J\ m^{-1}\ s^{-1}\ K^{-1}$
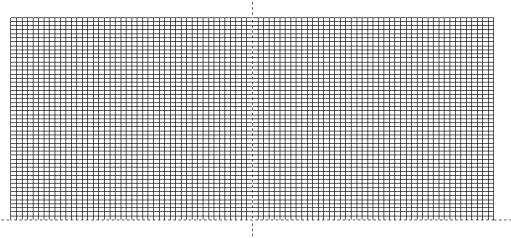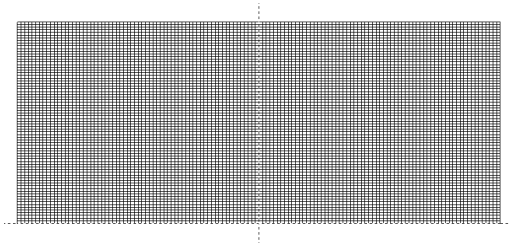
mesh A

mesh B

Figure 17.5 Finite element meshes used: mesh A consisting of 4400 element and 4539 nodes, refined mesh B with 9900 elements and 10108 nodes.

The finite element meshes as shown in Fig. 17.5 which have proven to be capable of attaining convergent solutions for the Elder problem[43] are also used for following investigations.

17.7.2.2 Results and discussion

The basis for comparison is the thermohaline simulation for the pure saline free convection, i.e. $Ra_t = 0$ and $B = \infty$, as presented in the first part of this paper[43]. It meets the best numerical approximation available for this case: divergent formulation of the mass transport equation, extended Boussinesq approximation, Galerkin-FEM, and predictor-corrector AB/TR time integrator. As the convergence tolerance γ a value of 10^{-3} is used both for head h, salinity C and temperature T based on a RMS error norm (*cf.* eqn (17-21)).

To study the growing influence of thermohaline convection more in detail we consider the computational results using mesh A for decreasing buoyancy ratios $B = \infty$, 5, 4, 3, 2 as exhibited in a series of Fig. 17.6. While the results for $B = 5$ (Fig. 17.6b) are still rather similar to the pure (asymptotic) saline convection at $B = \infty$ (Fig. 17.6a), beginning with $B = 4$ the influence of the superimposing thermal convection on the salinity distribution becomes apparent (Figs. 17.6c-e). There are no more monotonic changes in the salinity pattern. Surprisingly, salinity distributions reveal asymmetric characteristics at longer times when the influence of thermal convection becomes stronger as seen at $B = 2$ in Fig. 17.6e.

To check the influence of spatial resolution the computations are repeated with the refined mesh B. The long-term salinity pattern for small buoyancy ratios are illustrated in Fig. 17.7. Now, symmetric salinity distributions appear for $B = 4$ (Fig. 17.7a) and $B = 2$ (Fig. 17.7b). A comparison with the coarser mesh counterparts of Fig. 17.6 reveals further qualitative changes in the pattern evolution. The case with an equilibrium of solutal and thermal buoyancy effects for $B = 1$ (Fig. 17.7c) gives again asymmetric distributions of salinity. Note, the effective Rayleigh number is here

already 800 ($Ra = Ra_s + Ra_t$) where transient disturbances should take influence. However, there is apparently no physical reason for a broken symmetry and mesh effects are likely responsible for such an asymmetric evolution. It is obvious, at sufficiently high Rayleigh numbers each initially small disturbance which is not perfectly symmetric can evoke asymmetry which grows over a longer period. Moreover, in the numerical solution process such disturbances can be caused, e.g., by inappropriate spatial discretizations, remaining errors in solving the matrix systems by iterative techniques or roundoff errors arising in computing the physically instable process. On the other hand, in a physical experiment or in real sites the trigger of asymmetry may be an initially disturbed distribution or due to nonhomogeneous materials.

It seems that the numerical solutions reflect the physical instabilities which is most apparent for the thermohaline system if the solutal and thermal effects are nearly equilibrated ($B = 1$). It becomes obvious that modeling of such unstable thermohaline systems will be very expensive, especially in 3D.

Finally, Fig. 17.8 presents both the simulated temperature and salinity distributions for the case of $B = 4$. It demonstrates how the salinity evolution in a thermohaline convection process is related to specific pattern formations of the temperature field.
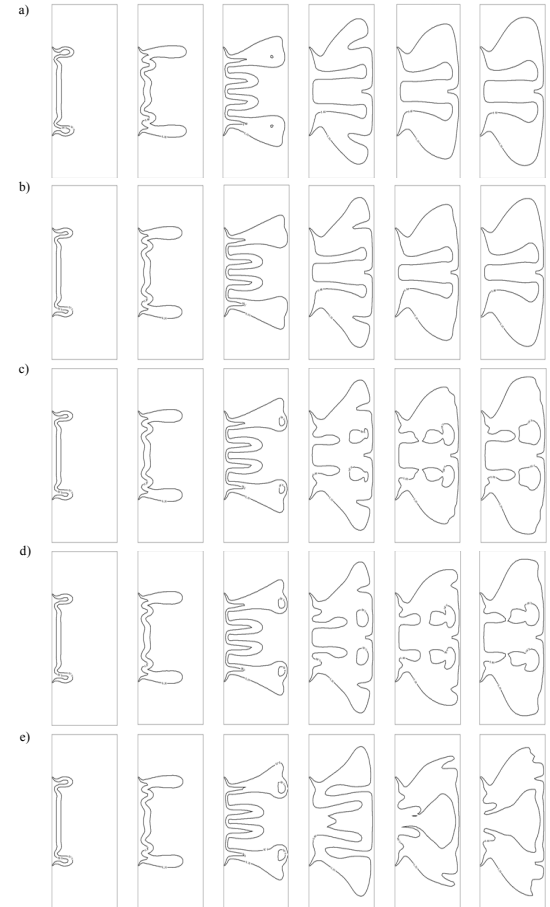
Figure 17.6 Influence of thermohaline convection: computed salinity distributions of 0.2 and 0.6 normalized isochlors at 1, 2, 4, 10, 15, and 20 years (from left to right) for different buoyancy ratios (a) $B = \infty$, (b) $B = 5$, (c) $B = 4$, (d) $B = 3$, and (e) $B = 2$ by using mesh A.

Figure 17.8 Computed distributions of salinity and temperature at several times for $B = 4$ using mesh A.

Figure 17.7 Mesh effects: computed salinity distributions of 0.2 and 0.6 normalized isochlors at 10, 15, and 20 years (from left to right) for different buoyancy ratios (a) $B = 4$, (b) $B = 2$, and (c) $B = 1$ by using mesh B.

17.7.3 The 2D thermohaline salt dome problem

17.7.3.1 Definition of the problem

The considered test case is an idealization of the flow over a salt dome[32,43,44,55], where the geometry is greatly simplified. The geometry and boundary conditions used are shown in Fig. 17.9. The cross section of the model extends horizontally 900 m and vertically 300 m having an aspect ratio A of 3. The aquifer is considered to be homogeneous and isotropic. The hydraulic head varies linearly on the top of the aquifer. All remaining boundaries are impervious to flow. The salinity on the top is taken equal to zero (freshwater) over the entire boundary. Additionally, a minimum mass flux constraint condition of $Q_C^{min_1} \equiv 0$ is imposed. It controls that the freshwater condition is only valid if the flow enters the domain. The middle section of the aquifer base represents the cap of the salt dome having a relative salt concentration equal to unity. The thermohaline extension of the salt dome problem concerns a superimposition of a thermal gradient acting upward and it tends to destabilize the brine pool due to the arising buoyant forces. Accordingly, the bottom of the aquifer is assigned by a constant normalized temperature of $T = 1$, while the top boundary is imposed by a normalized temperature of zero ($T = 0$). Again, the upper boundary is additionally constrained by a minimum heat flux of zero $Q_T^{min_1} \equiv 0$ which permits a control of the boundary conditions for inflowing and outflowing situations. The side walls of the domain are regarded as impervious for solute mass and adiabatic (insulated) for heat. The model parameters are summarized in Tab. 17.2. According to the DDC classification as stated above, the formulation of the thermohaline salt dome problem is one of a diffusive regime where the buoyancy force is caused by heat, which has a larger diffusivity than salt.
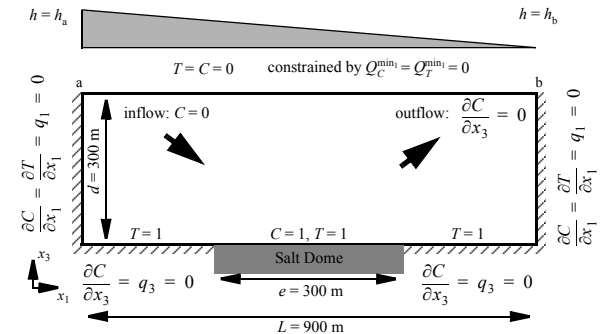
Figure 17.9 Definition of the 2D thermohaline salt dome problem (modified from Herbert *et al.*[32]).

Table 17.2 Simulation parameters for the 2D thermohaline salt dome problem

Symbol	Quantity	Value	Unit
A	aspect ratio	3	1
B	buoyancy ratio (Turner number)	2, 3, 5	1
C_0	reference concentration	0.	g l^{-1}
$c^f \rho^f$	thermal capacity of fluid	$4.2 \cdot 10^6$	J m^{-3} K^{-1}
$c^s \rho^s$	thermal capacity of solid	$2.52 \cdot 10^6$	J m^{-3} K^{-1}

Table 17.2 Simulation parameters for the 2D thermohaline salt dome problem (continued)

Symbol	Quantity	Value	Unit
D_d	molecular diffusion coefficient	$1.39 \cdot 10^{-8}$	$m^2\,s^{-1}$
d	thickness (height)	300	m
e	extent of intrusion	300	m
f_μ	viscosity relation function	1	1
h_a	hydraulic head at point a	10.228	m
h_b	hydraulic head at point b	0.	m
K	hydraulic conductivity	$1.0985252 \cdot 10^{-5}$	$m\,s^{-1}$
Le	Lewis number	217	1
Ra_s	solutal Rayleigh number	$2.4 \cdot 10^5$	1
Ra_t	thermal Rayleigh number	547, 365, 219	1
T_0	reference temperature	1.	K
α_L	longitudinal thermodispersivity	20.	m
α_T	transverse thermodispersivity	2.	m
$\bar{\alpha}/C_s$	density ratio	0.2036108	1
β_L	longitudinal dispersivity of solute	20.	m
β_T	transverse dispersivity of solute	2.	m

Table 17.2 Simulation parameters for the 2D thermohaline salt dome problem (continued)

Symbol	Quantity	Value	Unit
$\bar{\bar{\beta}}$	thermal expansion coefficient	$5 \cdot 10^{-4}$	K^{-1}
ε	porosity	0.2	1
Λ	thermal diffusivity	$6.024 \cdot 10^{-7}$	$m^2\,s^{-1}$
λ^f	thermal conductivity of fluid	0.65	$J\,m^{-1}\,s^{-1}\,K^{-1}$
λ^s	thermal conductivity of solid	3.	$J\,m^{-1}\,s^{-1}\,K^{-1}$

The finite element mesh as shown in Fig. 17.10 is used for the simulations of the thermohaline salt dome problem. The predictor-corrector AB/TR time integrator with a RMS-based convergence tolerance γ of 10^{-3} is applied.
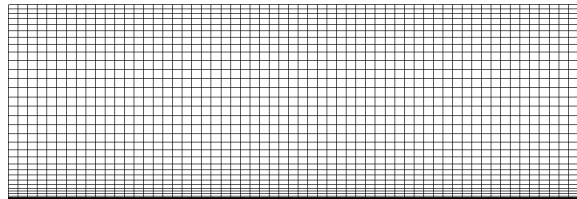
Figure 17.10 Finite element mesh used for 2D thermohaline salt dome problem consisting of 1920 elements and 2013 nodes.

17.7.3.2 Results and discussion

Simulated results of the salt dome problem at a time of 100 years for different buoyancy ratios B are shown in Fig. 17.11. It reveals the temperature effect on the saltwater distribution remains negligible or small if compared with the single-diffusive results[43] at higher buoyancy ratios B. As seen for $B = 2$, however, if the buoyancy ratio becomes smaller vigorous temperature influences on the brine pattern result in form of a 'wavy' salinity field caused by the thermal buoyancy.
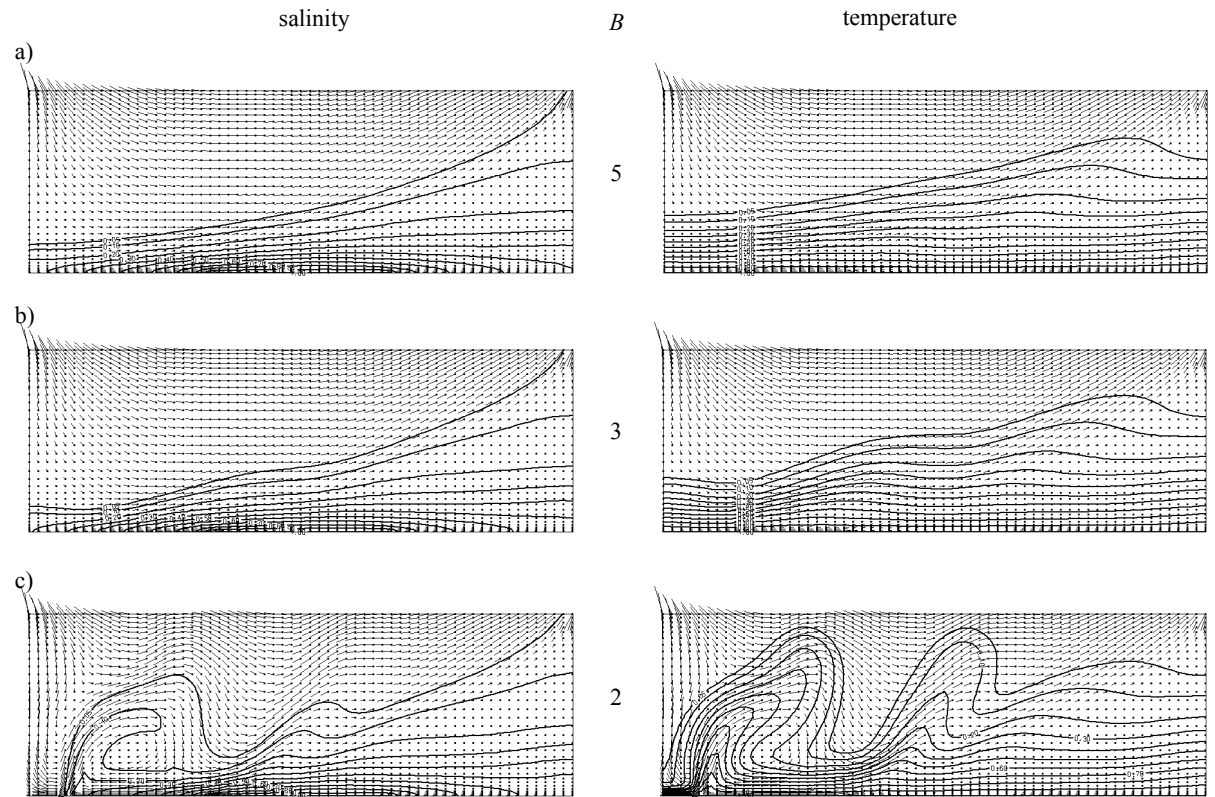
Figure 17.11 Evolution of the thermohaline convection system: computed salinity and temperature distributions at 100 years for different buoyancy ratios (a) $B = 5$, (b) $B = 3$, and (c) $B = 2$.

To illustrate how such a thermal effect on the brine flow is evolved a series of salinity and temperature patterns are outlined in Fig. 17.12 for the case of $B = 2$. The 'wavy' salinity characteristics is triggered in front of the salt wedge by thermally driven eddies. As expected, it leads to an increased saltwater effluent on top of the aquifer. Note, a buoyancy ratio of 2 implies an already large temperature difference for a high-concentration brine and, accordingly, corresponds to an extreme situation. It should be mentioned that for the real site behind the present salt dome problem such high temperatures corresponding to $B = 2$ may be unlikely to occur in practice. However, the variants can be valuable as test cases to study the effects of higher temperatures, which may, for instance, arise in the vicinity of a disposal facility for heat-emitting waste.
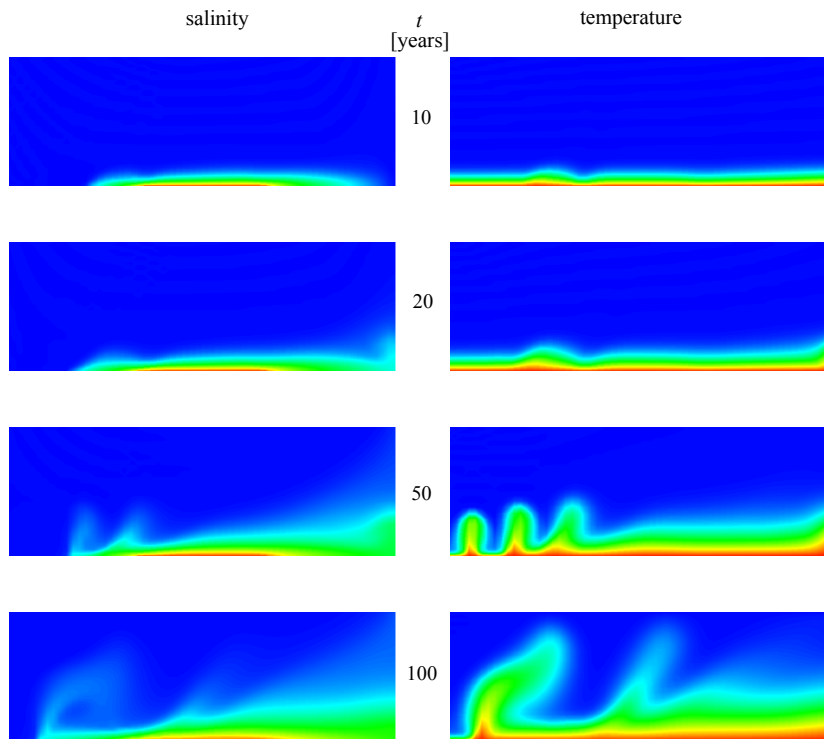
Figure 17.12 Evolution of the thermohaline convection system: computed salinity and temperature distributions at several times for a buoyancy ratio of $B = 2$.

17.8 Examples of 3D Cellular Convection

17.8.1 The 3D Elder problem for single-diffusive (solutal) and double-diffusive (thermohaline) convection

17.8.1.1 Definition of the problem

Originally, the Elder problem[23] refers to a 2D cross-sectional convection process in a fluid-saturated porous layer. As a result, only 2D roll cells can appear. Now, the interest is focussed on adequate 3D situations. For this purpose the Elder problem is expanded for both the single-diffusive and double-diffusive applications in a porous box consisting of a square base ($L \times L$) and a height d. This box has the same cross sections along the Cartesian axes as defined in Fig. 17.4 for the 2D sketch. Boundary conditions and measures are identical to the 2D case shown in Fig. 17.4. Now, salinity is held constant in an areal extent on top and bottom of the porous box. The used parameters correspond to those given in Tab. 17.1.

The box is discretized by hexahedral trilinear finite elements as displayed in Fig. 17.13. To reduce the computation effort only a quarter of the discretized domain is actually simulated. It is based on the assumption that symmetric planes occur for the studied range of Rayleigh numbers. Both AB/TR and FE/BR time marching with a RMS-based convergence tolerance γ of 10^{-3} have been tested. For the long-term simulations and the chosen spatial resolution the second-order AB/TR scheme with a full Newton method becomes sensitive and produces oscillations at later simulation times. On the other hand, the first-order FE/BE scheme with full Newton method has proven to be more stable and robust and, therefore, it is preferred for present 3D simulations. Generally, Galerkin-FEM (i.e. no upwinding) is used. To simulate the convection process over a period of 100 years the FE/BE scheme takes 641 time steps for the single-diffusive problem and 965 time steps for the double-diffusive (thermohaline) problem (excluding restarted steps).

simulated mesh quarter
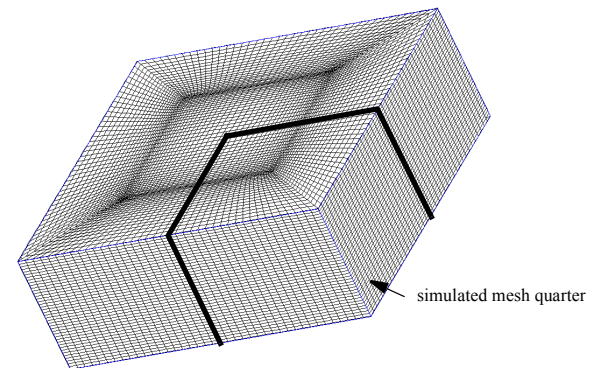
Figure 17.13 Total finite element mesh for the 3D Elder problem: only a quarter of the mesh is actually used in the computation. This quarter consists of 48,000 hexahedral elements and 51,701 nodes.

17.8.1.2 Results and discussion

The 3D free convection process is similar to the 2D counterpart, with some interesting new features. To give more insight into the physics of the 3D convection

process Fig. 17.14 shows the evolution of salinity from different views. The 3D cut-away images (left column of Fig. 17.14) display the progressing fingering characteristics in the 3D space. Similar to the 2D case we find also an upwelling salinity pattern in the centre of the box at the given time stages. The 3D influence becomes also apparent in the two horizontal views at an upper elevation of $0.9 \cdot d$ (135 m) and the middle horizon of $0.5 \cdot d$ (75 m) as shown in Fig. 17.14. At the beginning the quadratic geometry of the intrusion area on top is visible in the convection pattern. Fingers appear around the border of the intrusion area and 'blobs' grow down at the four corners. The quadratic pattern evolves into more complicated multicellular formations via a number of characteristic stages. More 'blobs' appear up to the time when the salinity reaches the bottom. Then, the structures begin to fuse and the pattern is completely reformed. After this phase a convection pattern remains which has a characteristic diagonal 'star' form. This 'star' is a result of the geometry of the square intrusion area. It becomes clear that the final formations have a strong dependency on the geometric relations.

An illustration of the pattern evolution in 3D space is given in Fig. 17.15 where isosurfaces of the 50% salinity are shown at characteristic time stages. Up to a time of about 4 years the salinity primarily sinks down and forms a dissected finger formation. At later time the upper part contracts and forms the typical diagonal 'star', while larger 'blobs' are getting fused below.
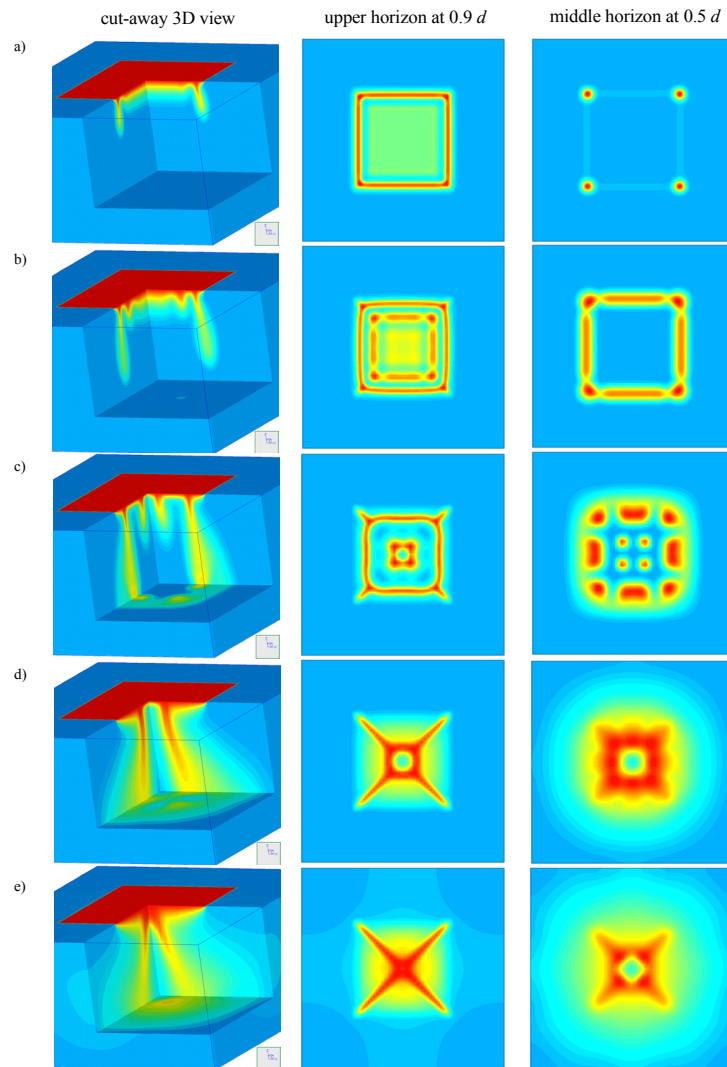
Figure 17.14 Computed salinity patterns of the 3D Elder problem at times of (a) 1, (b) 2, (c) 4, (d) 10, and (e) 20 years.
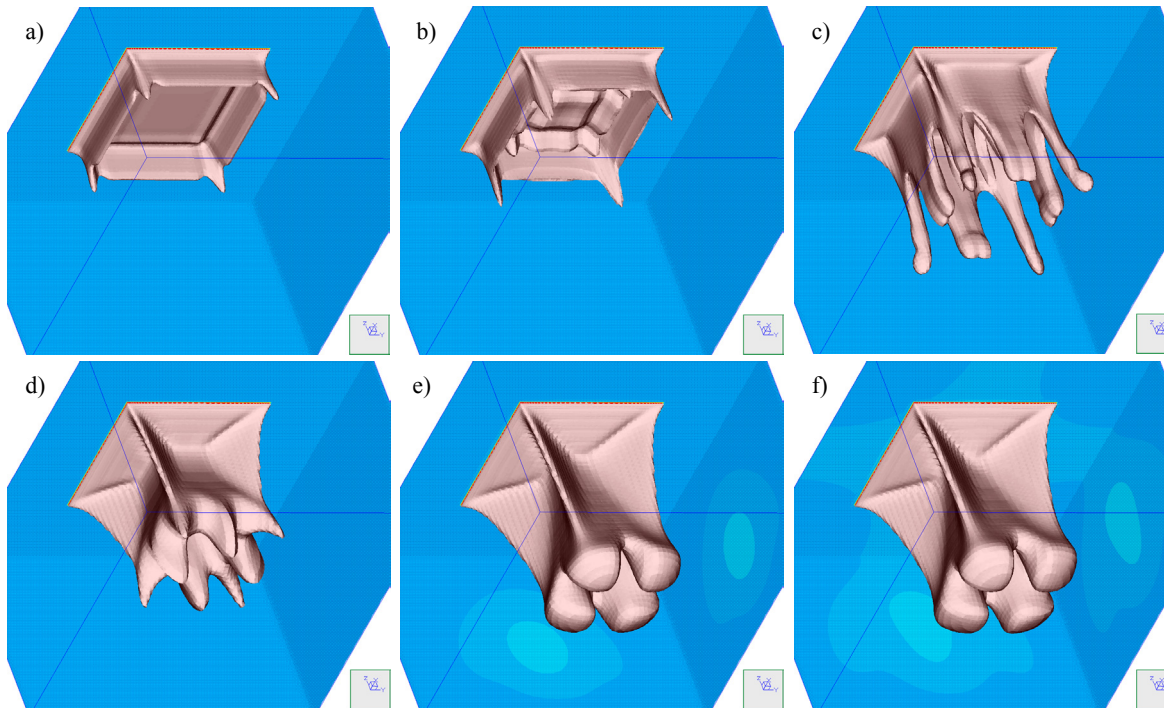
Figure 17.15 Computed 3D isosurfaces of 50% salinity for the 3D Elder problem (viewing into the box from bottom to top) at times of (a) 1, (b) 2, (c) 4, (d) 10, (e) 15, and (f) 20 years.

The 3D thermohaline Elder problem has been simulated for a buoyancy ratio of $B = 5$, where the solutal Rayleigh number Ra_s is again 400. The 3D distributions of the computed salinities and temperatures up to 20 years are displayed in Fig. 17.16. In contrast to the single-diffusive formation (*cf.* Fig. 17.14) the salinity pattern appears more diffusive at later times when the temperature field affects the convection system. Then, the thermally buoyant forces accelerate the contraction process of the sinking salinity plume in the centre. At the final stage, while the single-diffusive convection provides still an upwelling flow in the centre, the thermohaline convection process reveals a single downwelling characteristics for the salinity (see Figs. 17.15 and 17.17). As seen, the most heated water is buoyantly affected outside and around the denser salinity core, where the isotherms come to the upper locations. These mutual influences between salinity and temperature are more apparent in Figs. 17.17 and 17.18 for the computed isosurfaces of salinity and temperature, respectively.
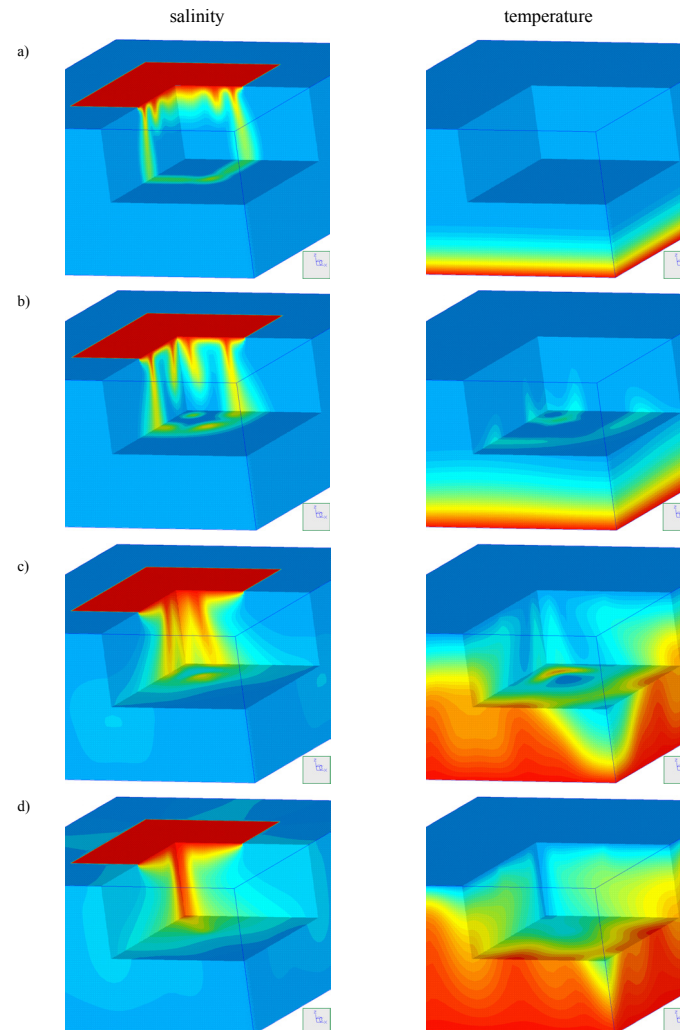
Figure 17.16 Cut-away views of simulated salinity (left) and temperature (right) distributions for the 3D thermohaline Elder problem at buoyancy ratio of $B = 5$ and times of (a) 2, (b) 4, (c) 10, and (d) 20 years.
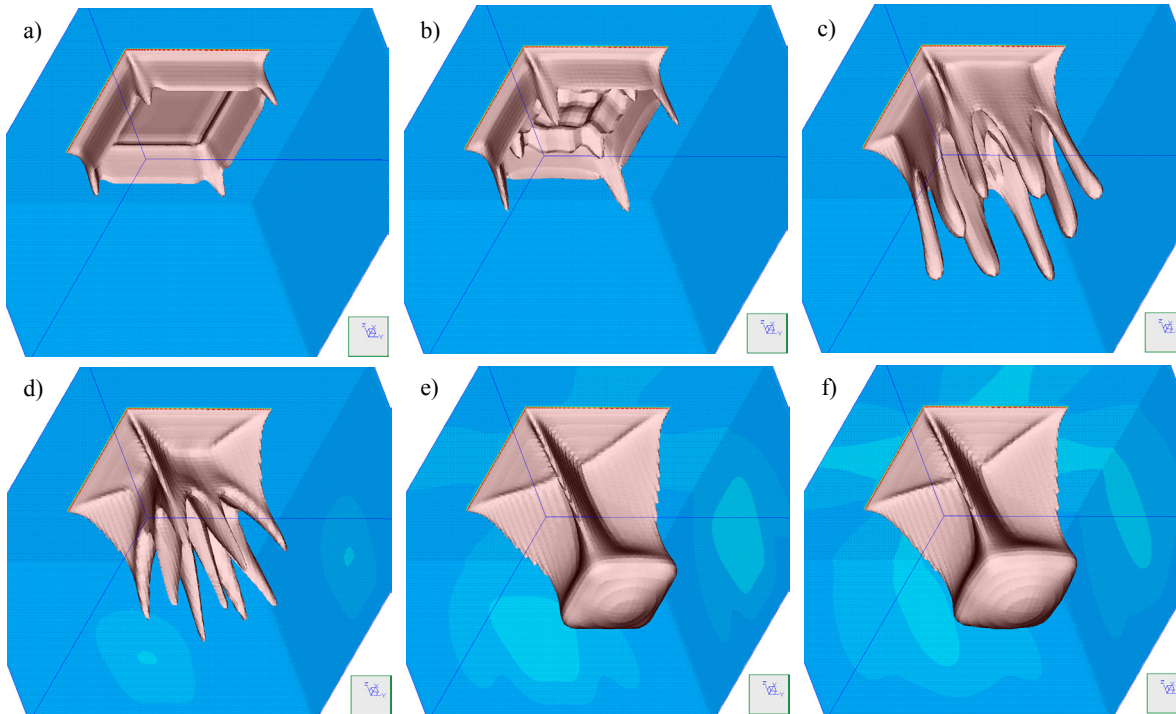
Figure 17.17 Computed 3D isosurfaces of 50% salinity for the 3D thermohaline Elder problem (viewing from bottom to top) at $B = 5$ and times of (a) 1, (b) 2, (c) 4, (d) 10, (e) 15, and (f) 20 years.
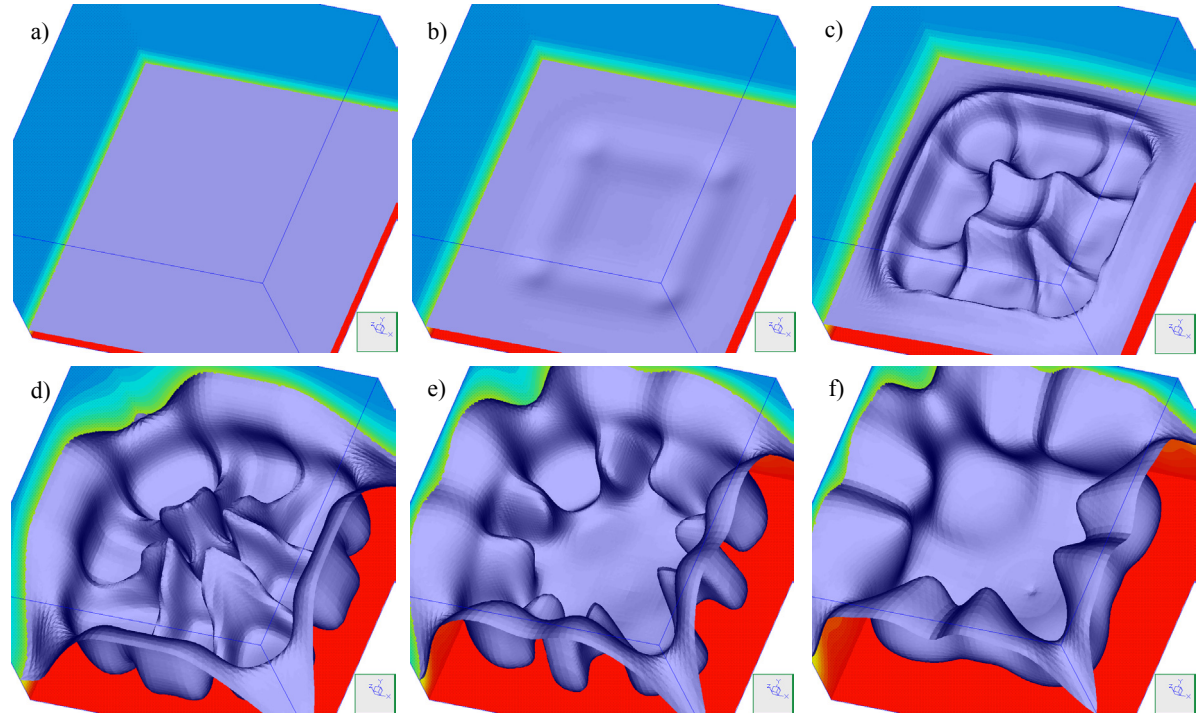
Figure 17.18 Computed 3D isosurfaces of 50% temperature for the 3D thermohaline Elder problem (viewing from top to bottom) at $B = 5$ and times of (a) 1, (b) 2, (c) 4, (d) 10, (e) 15, and (f) 20 years.

17.8.2 The 3D Bénard convection

17.8.2.1 Definition of the problem

Three-dimensional convective pattern formations in domains representing a thin porous layer, i.e., for large aspect ratios A, can be considered as a porous medium equivalent of Bénard convection. As Elder[23] studied such a problem in 2D referred to as the 'long-heater problem' for a Rayleigh number of 200, an aspect ratio A of 10, and an intrusion ratio I of 0.8. We extend this 'long-heater problem' to 3D similar to the above Elder problem. The remaining simulation parameters correspond to that of the original Elder problem described in the first part of this paper[43]. Due to the multicellular

convection process in the porous layer a more refined spatial discretization is needed compared with the 3D Elder problem above. Moreover, no assumptions of symmetry are made and, accordingly, the domain has to be fully discretized. The finite element mesh for the problem consists of 220,000 (100 x 100 x 22) hexahedral trilinear elements containing 234,623 (101 x 101 x 23) nodes. Again, for the temporal discretization the FE/BE predictor-corrector scheme with the full Newton method and a RMS-based convergence tolerance γ of 10^{-3} is applied to the simulation.

17.8.2.2 Results and discussion

The striking features of 3D Bénard convection development are shown in Fig. 17.19. The initial motion is characterized by a rectangular string of end-cells, where at the four corner points the most intensive growths of 'blobs' can be observed. It is followed by a growth of cells starting from the ends of the intrusion area on top. At these times a remarkable feature of the 3D convection process is the annular roll pattern formation. At smaller times the cell structures are rather complex (Fig. 17.19b) showing the birth of subcellular eddies both across and along the annular structure. Due to the smaller Rayleigh number the nonroll-like perturbations are smoothed at larger times and the convection process results in a highly regular pattern of ring structures.

water systems. The described solution strategies as implemented in the simulator FEFLOW are more general and are primarily developed to tackle complex practical applications where solutal and/or thermal density effects play an important role. However, before more complex field situations can be studied the chosen methods and codings have to be extensively tested over a wider spectrum of this important class of nonlinear problems. In this context the aim of the present paper is mainly the proving and benchmarking of the simulations along examples where comparable results are available, or if not, the obtained results are to be supposed as a comparison basis for further studies. We have chosen the Elder and salt dome problem (HYDROCOIN case 5 level 1) as well suited and representative examples. They allow us both to participate in the process of resolving partly contradictory results given in the literature and to expand (or generalize) the 2D solutions to three dimensions and additional coupling phenomena from a well-documented and accepted source. The extensions concern thermohaline and multicellular convection processes in 2D and 3D. Unfortunately, to date both numerical and experimental results of 3D and thermohaline convection are rare and we are mostly dependent on an incremental procedure in comparing and interpreting the results among one another. In this context we found similarities and also interesting new features regarding the pattern formations of the buoyancy-driven convection processes.

17.9 Closure

The finite-element method is applied to simulate variable density flow processes in 2D and 3D ground-

vertical distributions at fences   50% salinity isosurface   horizon at 0.9 d

viewing from bottom to top
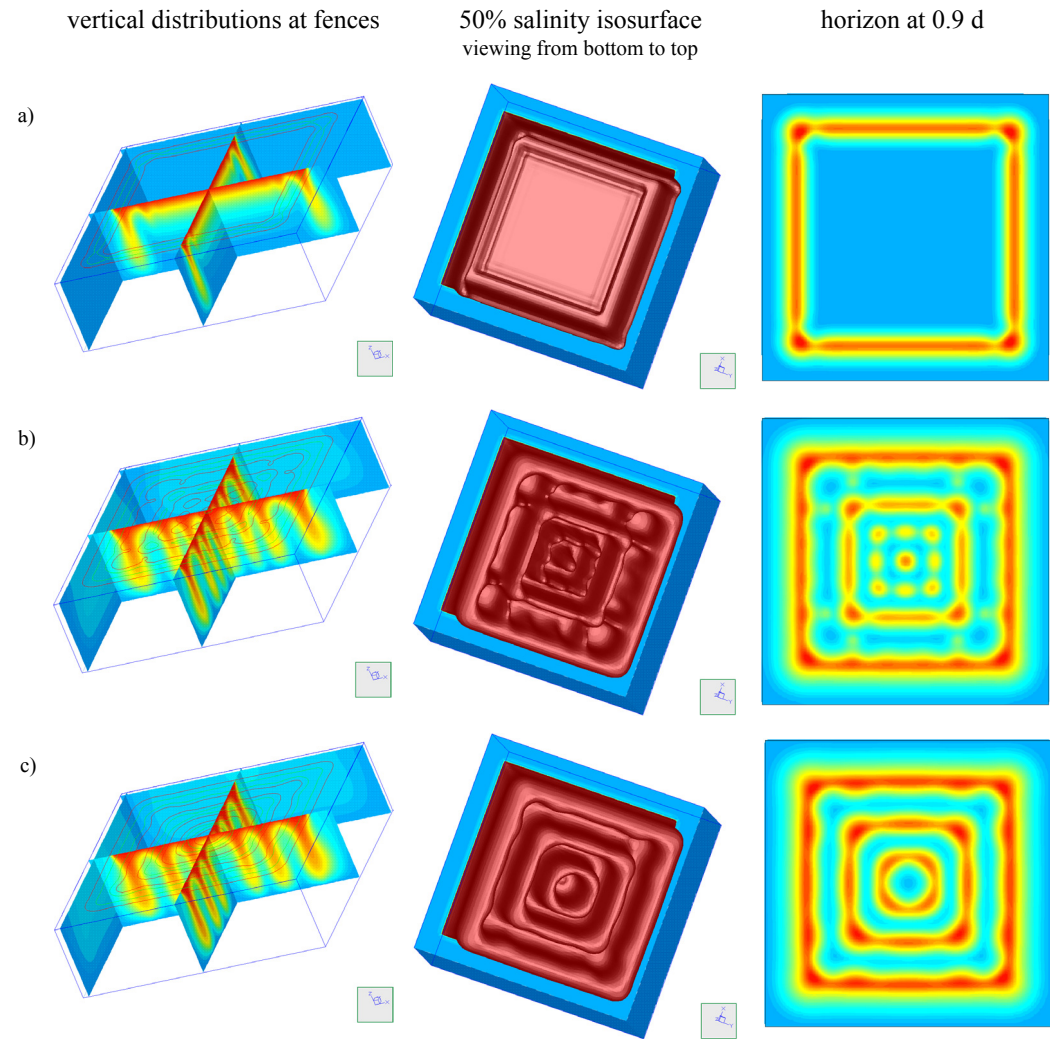
Figure 17.19 Computed salinity patterns for the 3D Bénard convection problem at Rayleigh number of 200 and dimensionless times of (a) 0.013, (b) 0.026, and (c) 0.078.

The interaction between solutal and thermal convection is studied by varying the buoyancy ratio B, which expresses the relationship between buoyancy forces due to solutal and thermal convection. Differences between (pure) saline convection and thermohaline convection become apparent for buoyancy ratios $B \le 5$. We found asymmetric convection patterns for buoyancy ratios near to unity. In this situation, the hydrodynamic system becomes strongly unstable because the solutal and thermal buoyancy effects are nearly equilibrated. As a result, very small vertical velocities trigger the convection process. Grid effects indicate the physical instability. The numerical solution of thermohaline convection systems with buoyancy ratios near to unity requires extremely fine spatial discretizations.

Three-dimensional convection needs sufficiently high spatial and temporal resolutions if damping measures, such as upwinding, are to be avoided. At moderate Rayleigh numbers (400 for the 3D Elder problem and 200 for the 3D Bénard convection) we used more than 50,000 nodes for a quarter of the domain and about 230,000 nodes for the total discretization of a 3D porous layer subjected to a free convection process. In comparison with 2D, where it has been found[43] about 10,000 nodes are required to accomplish satisfactorily accurate results for the Elder problem, the chosen 3D resolution seems to be a minimum for this class of problems. Time marching is based on a predictor-corrector strategy with an automatic time step control embedded in a one-step full Newton method. For the present examples more than 600 intrinsic time steps are required for simulating a 3D convection process with a duration of about 20 years for a convergence tolerance γ of 10^{-3}.

It becomes clear that a long-term analysis of 3D free or thermohaline convection takes a large numerical effort and is normally a time-consuming task. While a 2D simulation is still on the order of hours of CPU time, a 3D problem can take days of runtime on a workstation. However, by using a high-speed workstation available today the 3D Elder problems and the Bénard convection could be solved in one day and two days of runtime, respectively. It should be taken into consideration that the FEFLOW code is general and not streamlined, for instance, for special cases of free convection in rectangular domains with homogeneous parameters. That means the impact and the found efforts are representative for general problems having an arbitrary geometry and permitting such parametric and boundary conditions which are required in actual site-specific applications[19].

References

1. Ackerer, P., Mosé, R., Siegel, P. & Chavent, G., Reply to contribution of reference 11. *Water Resour. Res.*, 32 (1996) 1911-1913.

2. Angirasa, D. & Scrinivasan, J., Natural convection flows due to the combined buoyancy of heat and mass diffusion in a thermally stratified medium. *J. Heat Transfer*, **111** (1989) 657-663.

3. Axelsson, O., *Iterative solution methods*, Cambridge Univ. Press, 1994.

4. Beauwens, R., Modified incomplete factorization strategies. In: *Preconditioned Conjugate Gradient Methods, Lecture Notes in Math. 1457* (ed. by O. Axelsson & L. Kolotilina), Springer, New York, 1990, 1-16.

5. Behie, A. & Vinsome, P. K. W., Block iterative methods for fully implicit reservoir simulation. *SPEJ Soc. Pet. Eng. J.,* (October 1982) 658-668.

6. Beukema, K. J. & Bruin, S., Three-dimensional natural convection in a confined porous medium with internal heat generation. *Int. J. Heat Mass Transfer,* **26** (1983) 451-458.

7. Bixler, N. E., An improved time integrator for finite element analysis. *Communications in Applied Numerical Methods,* **5** (1989) 69-78.

8. Brandt, A. & Fernando, H. J. S. (Eds.), *Double-diffusive convection*, Geophysical Monograph Vol. 94, American Geophysical Union, Washington, D.C., 1995.

9. Caltagirone, J. P., Meyer, G. & Mojtabi, A., Structurations thermoconvectives tridimensionnelles dans une couche poreuse horizontale. *J. Mécanique,* **20** (1981) 219-232.

10. Chan, Y. T. & Banerjee, S., Analysis of transient three-dimensional natural convection in porous media. *ASME J. Heat Transfer,* **103** (1981) 242-248.

11. Cordes, Ch. & Kinzelbach, W., Comment on contribution of reference 51. *Water Resour. Res.*, **32** (1996) 1905-1909.

12. Croucher, A. E. & O'Sullivan, M. J., The Henry problem for saltwater intrusion. *Water Resour. Res.,* **31** (1995) 1809-1814.

13. Diersch, H.-J., Primitive variables finite element solutions of free convection flows in porous media. *Zeitschr. Angew. Math. Mech. (ZAMM),* **61** (1981) 325-337.

14. Diersch, H.-J., Study of free convective flows in porous media and effects of mechanical dispersion using finite element methods. *Gerlands Beiträge zur Geophysik,* **90** (1981) 489-506.

15. Diersch, H.-J., On finite element upwinding and its numerical performance in simulating coupled convective transport processes. *Zeitschr. Angew. Math. Mech. (ZAMM),* **63** (1983) 479-488.

16. Diersch, H.-J., Prochnow, D. & Thiele, M., Finite element analysis of dispersion-affected saltwater upconing below a pumping well. *Appl. Math. Modelling,* **8** (1984) 305-312.

17. Diersch, H.-J., Finite element modelling of recirculating density-driven saltwater intrusion processes in groundwater. *Adv. Water Resour.,* **11** (1988) 25-43.

18. Diersch, H.-J. G., Computational aspects in developing an interactive 3D groundwater transport simulator using FEM and GIS. *Groundwater Quality Management* (Proc. of the GQM 93 Conf. held at Tallinn, Sept. 1993). IAHS Publ. no. 220, 1994, 313-326.

19. Diersch, H.-J., Albert, H., Schreyer, J. & Richter, J., Three-dimensional modeling of flow and contaminant transport processes arising in flooding the Königstein uranium pit. In: B. Merkel *et al.* (eds): *Proc. Intern. Conf. Uranium-Mining and Hydrogeology*, Verlag Sven von Loga, Köln, 1995, 121-130.

20. Diersch, H.-J. G., Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow, contaminant mass and heat transport processes. WASY Ltd., Berlin, 2002.

21. Diersch, H.-J. G. & 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 held at Golden, Colorado, September 1996), IAHS Publ. no. 237, 1996, 407-415.

22. Diersch, H.-J. G., FEFLOW - Physical basis of modeling, Reference Manual - Part I, WASY Ltd., Berlin, 2002.

23. Elder, J. W., Transient convection in a porous medium. *J. Fluid Mech.*, **27**, Part 3 (1967) 609-623.

24. Evans, G. E. & Nunn, J. A., Free thermohaline convection in sediments surrounding a salt column. *J. Geophysical Research*, **94** (1989) 12413-12422.

25. Fan, Y. & Kahawita, R., A numerical study of variable density flow and mixing in porous media. *Water Resour. Res.,* **30** (1994) 2707-2716.

26. Frind, E. O., Simulation of long-term transient density-dependent transport in groundwater. *Adv. Water Resour.*, **5** (1982) 73-88.

27. Galeati, G., Gambolati, G. & Neuman, S. P., Coupled and partially coupled Eulerian-Lagrangian model of freshwater-seawater mixing. *Water Resour. Res.,* **28** (1992) 149-165.

28. Gartling, D. K. & Hickox, C. E., Numerical study of the application of the Boussinesq approximation for a fluid-saturated porous medium. *Int. J. Numer. Methods Fluids,* **5** (1985) 995-1013.

29. Gresho, P. M., Lee, R. L., Chan, S. T. & Leone Jr., J. M., A new finite element for Boussinesq fluids. *Preprint UCRL-82842*, Lawrence Livermore Lab., Univ. of California, 1979, 12p.

30. Gresho, P. M., Lee, R. L. & Sani, R.L., On the time-dependent solution of the incompressible Navier-Stokes equations in two and three dimensions. *Preprint UCRL-83282*, Lawrence Livermore Lab., Univ. of California, 1979, 53p.

31. Hassanizadeh, M. S. & Leijnse, A., On the modeling of brine transport in porous media. *Water Resour. Res.,* **24** (1988) 321-330.

32. Herbert., A. W., Jackson, C. P. & Lever, D. A., Coupled groundwater flow and solute transport with fluid density strongly depen-

dent upon concentration. *Water Resour. Res.,* **24** (1988) 1781-1795.

33. Hestenes, M. R. & Stiefel, E., Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand. Sect. B,* **49** (1952) 409-436.

34. Holst, P. H. & Aziz, K., Transient three-dimensional natural convection in confined porous media, *Int. J. Heat Mass Transfer.,* **15** (1972) 73-90.

35. Horne, R. N. & O'Sullivan, M. J., Oscillatory convection in a porous medium heated from below. *J. Fluid Mech.,* **66** (1974) 339-352.

36. Horne, R.N., Three-dimensional natural convection in a confined porous medium heated from below. *J. Fluid Mech.,* **92** (1979) Part 4, 751-766.

37. Horne, R. N. & Caltagirone, J. P., On the evaluation of thermal disturbances during natural convection in a porous medium. *J. Fluid Mech.,* **100** (1980) 385-395.

38. Huyakorn, P. S. & Taylor, C., Finite element models for coupled groundwater and convective dispersion. *Proc. 1st Int. Conf. Finite Elements in Water Resources,* (ed. Gray, W. G. *et al.*) Princeton Univ., July 1976, Pentech Press, London, 1977, 1.131-1.151.

39. Huyakorn, P. S., Anderson, P. F., Mercer, J. W. & White Jr., H. O., Saltwater intrusion in aquifers: development and testing of a three-dimensional finite element model. *Water Resour. Res.,* **23** (1987) 293-312.

40. Kakinuma, T., Kishi, Y. & Inouchi, K., The behaviour of groundwater with dispersion in coastal aquifers. *J. Hydrology,* **98** (1988) 225-248.

41. Knabner, P. & Frolkovic, P., Consistent velocity approximation for finite volume or element discretization of density driven flow in porous media. *Proc. XI. Int. Conf. Computational Methods in Water Resources* held in Cancun, México, July 1996, (ed. Aldama *et al.*) Comp. Mech. Publ., Vol. 1, Southampton, 1996, 93-100.

42. Kolditz, O., Benchmarks for numerical groundwater simulations, In: *Diersch, H.-J. G.: FEFLOW User's Manual, Chapter 5,* WASY Ltd., Berlin, 1996.

43. Kolditz, O., Ratke, R., Diersch, H.-J. G. & Zielke, W., Coupled groundwater flow and transport: 1. Verification of variable-density flow and transport models, *Adv. Water Resour.,* this issue.

44. Konikow, L. F., Campbell, P. J. & Sanford, W. E., Modelling brine transport in a porous medium: a re-evaluation of the HYDROCOIN Level 1, Case 5 problem. *Calibration and Reliability in Groundwater Modelling* (Proceedings of the Model-CARE 96 Conference held at Golden, Colorado, September 1996), IAHS Publ. no. 237, 1996, 363-372.

45. Lee, R. L., Gresho, P. M. & Sani, R. L., Numerical smoothing techniques applied to some finite element solutions of the Navier-Stokes equations. *Proc. 1st Int. Conf. Finite Elements in Water Resources,* (ed. Gray, W. G. *et al.*) Princeton Univ., July 1976, Pentech Press, London, 1977, 4.127-4.145.

46. Leijnse, A., Three-dimensional modeling of coupled flow and transport in porous media. *Dissertation,* University of Notre Dame, Indiana (USA), 1992.

47. Letniowski, F. W., An overview of preconditioned iterative methods for sparse matrix equations. Research Report CS-89-26, Faculty of Math., Univ. of Waterloo, Can., 1989, 32p.

48. Lever, D. A. & Jackson, C.P., On the equations for the flow of concentrated salt solution through a porous medium. U.K. DOE Report No. DOE/RW/85.100, 1985.

49. Meijerink, J. A. & Van der Vorst, H. A., Guidelines for the usage of incomplete decomposition in solving sets of linear systems as they occur in practical problems. *J. Comput. Phys.,* **44** (1981) 134-155.

50. Mercer, J. W. & Pinder, G.F., Finite element analysis of hydrothermal systems. *Finite Element Methods in Flow Problems (Proc. 1st Symp., Swansea, 1974),* (ed. Oden, J. T. *et al.*) Univ. of Alabama Press, 1974, 401-414.

51. Mosé, R., Siegel, P., Ackerer, P. & Chavent, G., Application of the mixed hybrid finite element approximation in a groundwater flow model: Luxury or necessity? *Water Resour. Res.,* **30** (1994) 3001-3012.

52. Murray, B. T. & Chen, C. F., Double-diffusive convection on a porous medium. *J. Fluid Mech.,* **201** (1989) 147-166.

53. Nield, D. A., Onset of thermohaline convection in a porous medium. *Water Resour. Res.,* **11** (1968) 553-560.

54. Nield, D. A. & Bejan, A., *Convection in porous media,* Springer Verlag, Berlin 1992.

55. Oldenburg, C. M. & Pruess, K., Dispersive transport dynamics in a strongly coupled groundwater-brine system. *Water Resour. Res.,* **31** (1995) 289-302.

56. Perrochet, P., Personal communication. EPFL Lausanne, GEOLEP Laboratoire de géologie, Lausanne (Switzerland), 1996.

57. Pinder, G. F. & Cooper, H. H., A numerical technique for calculating the transient position of the saltwater front. *Water Resour. Res.*, **6** (1970) 875-882.

58. Riley, D. S. & Winters, K. H., Time-periodic convection in porous media: the evolution of Hopf bifurcations with aspect ratio. *J. Fluid Mech.,* **223** (1991) 457-474.

59. Rubin, H., Onset of thermohaline convection in a cavernous aquifer. *Water Resour. Res.,* **12** (1976) 141-147.

60. Rubin, H. & Roth, C., Thermohaline convection in flowing groundwater. *Adv. Water Resour.,* **6** (1983) 146-156.

61. Saad, Y. & Schultz, M. H., GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.,* **7** (1986) 856-869.

62. Sauter, F. J. & Beusen, A. H. W., Streamline calculations using continuous and discontinuous velocity fields and several time integration methods. *Groundwater Quality Management* (Proc. of the GQM 93 Conf. held at Tallinn, Sept. 1993). IAHS Publ. no. 220, 1994, 347-355.

63. Schincariol, R. A., Schwartz, F. W. & Mendoza, C. A., On the generation of instabilities in variable density flow. *Water Resour. Res.,* **30** (1994) 913-927.

64. Schubert, G. & Straus, J.M., Three-dimensional and multicellular steady and unsteady convection in fluid-saturated porous media at high Rayleigh numbers. *J. Fluid Mech.,* **94** (1979) Part 1, 25-38.

65. Schubert, G. & Straus, J. M., Transitions in time-dependent thermal convection in fluid-saturated porous media. *J. Fluid Mech.,* **121** (1982) 301-303.

66. Segol, G., Pinder, G. F. & Gray, W.G., A Galerkin finite element technique for calculating the transient position of the saltwater front. *Water Resour. Res.*, **11** (1975) 343-347.

67. Shen, C. Y., The evolution of the double-diffusive instabilities: salt fingers. *Phys. Fluids*, **A1** (1989) 5.

68. Sonneveld, P., CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.,* **10** (1989) 36-52.

69. Straus, J.M. & Schubert, G., Three-dimensional convection in a cubic box of fluid-saturated porous material. *J. Fluid Mech.,* **91** (1979) Part 1, 155-165.

70. Straus, J. M. & Schubert, G., Modes of finite-amplitude three-dimensional convection in rectangular boxes of fluid-saturated material. *J. Fluid Mech.,* **103** (1981) 23-32.

71. Tong, C.H., A comparative study of preconditioned Lanczos methods for nonsymmetric linear systems. *Sandia Report SAND91-8240*, Sandia National Lab., Albuquerque, Sept. 1992, 118p.

72. Trevisan, O.V. and Bejan, A., Natural convection with combined heat and mass transfer buoyancy effects in a porous medium. *Int. J. Heat Mass Transfer,* **28** (1985) 1597-1611.

73. Tyvand, P. A., Thermohaline instability in anisotropic porous media. *Water Resour. Res.,* **16** (1980) 325-330.

74. Van der Vorst, H. A., Bi-CGSTAB: A fast and smoothly convergent variant of BiCG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.,* **13** (1992) 631-644.

75. Voss, C. I. & Souza, W. R., Variable density flow and solute transport simulation of regional aquifers containing a narrow freshwater-saltwater transition zone. *Water Resour. Res.,* **26** (1987) 2097-2106.

76. Xue, Y., Xie, C., Wu, J., Liu, P., Wang, J. & Jiang, Q., A three-dimensional miscible transport model for seawater intrusion in China. *Water Resour. Res.,* **31** (1995) 903-912.

77. Yeh, G. T., On the computation of Darcian velocity and mass balance in the finite element modeling of groundwater flow. *Water Resour. Res.,* **17** (1981) 1529-1534.

78. Zebib, A. & Kassoy, D. R., Three-dimensional natural convection motion in a confined porous medium. *Phys. Fluids,* **21** (1978) 1-3.

79. Zhang, H. & Schwartz, F. W., Multispecies contaminant plumes in variable density flow systems. *Water Resour. Res.,* **31** (1995) 837-847.

Appendix A

Weak form of the continuity equation (17-1)

The weak form of the continuity equation (17-1) gives

$$\int_\Omega wS_o \frac{\partial h}{\partial t} - \int_\Omega \frac{\partial w}{\partial x_i} q_i^f = -\int_\Gamma w q_i^f n_i + \int_\Omega w(Q_\rho + Q_{EB}) \quad \text{(A1)}$$

$$\frac{\rho^f - \rho_o^f}{\rho_o^f} = \frac{\bar{\alpha}}{(C_s - C_o)}(C - C_o) - \bar{\beta}(T - T_o) \quad \text{(A2)}$$

Introducing the Darcy equation (17-2) into eqn (17-1) and taking into account that the buoyancy term leads to

by using the fluid density equation of state (17-5), following final weighted residual formulation of the continuity equation results

$$\int_\Omega wS_o \frac{\partial h}{\partial t} + \int_\Omega \frac{\partial w}{\partial x_i} K_{ij} f_\mu \frac{\partial h}{\partial x_j} = -\int_\Omega \frac{\partial w}{\partial x_i} K_{ij} f_\mu \left[\frac{\bar{\alpha}}{(C_s - C_o)}(C - C_o) - \bar{\beta}(T - T_o) \right] e_j$$

$$-\int_\Omega w q_i^f \left(\frac{\bar{\alpha}}{(C_s - C_o)} \frac{\partial C}{\partial x_i} - \bar{\beta} \frac{\partial T}{\partial x_i} \right) - \int_\Omega w\varepsilon \left(\frac{\bar{\alpha}}{(C_s - C_o)} \frac{\partial C}{\partial t} - \bar{\beta} \frac{\partial T}{\partial t} \right) - \int_\Gamma w \underbrace{q_i^f n_i}_{q_{n_h}} + \int_\Omega wQ_\rho$$

$$\text{(A3)}$$

where the extended Boussinesq approximation term (17-7) is incorporated. Equation (A3) represents the standard weak form of the substitution formulation to solve the hydraulic head h.

Weak form of the Darcy equation (17-2)

Formally, a weak form of the Darcy equation can easily be derived as

$$\int_\Omega w q_i^f = -\int_\Omega wK_{ij} f_\mu \frac{\partial h}{\partial x_j} \quad \text{(A4)}$$

$$-\int_\Omega wK_{ij} f_\mu \left[\frac{\bar{\alpha}}{(C_s - C_o)}(C - C_o) - \bar{\beta}(T - T_o) \right] e_j$$

to solve the vector of Darcy fluxes q_i^f at given h, C and T.

Weak form of the mass transport equations (17-3)

The weak formulations for the divergent and convective forms of the contaminant mass conservation equations differ from the fact that for the former the divergence theorem is applied both to the convective and the dispersive terms

$$\int_\Omega w \frac{\partial}{\partial x_i}(q_i^f C - D_{ij} \frac{\partial C}{\partial x_j}) = -\int_\Omega \frac{\partial w}{\partial x_i} \left(q_i^f C - D_{ij} \frac{\partial C}{\partial x_j} \right) \quad \text{(A5)}$$

$$+ \int_\Gamma w \underbrace{\left(Cq_i^f - D_{ij} \frac{\partial C}{\partial x_j} \right) n_i}_{q_{n_C}^{total}}$$

while the conventional convective form applies the

divergence theorem only to the dispersive (2nd order) term

$$\int_\Omega w \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j}\right) = -\int_\Omega \frac{\partial w}{\partial x_i}D_{ij}\frac{\partial C}{\partial x_j} + \int_\Gamma w \underbrace{D_{ij}\frac{\partial C}{\partial x_j}n_i}_{-q_{n_C}^{disp}} \quad (A6)$$

Finally, it yields the following weak formulations

$$\int_\Omega wR_d\frac{\partial C}{\partial t} - \int_\Omega \frac{\partial w}{\partial x_i}q_i^f C + \int_\Omega \frac{\partial w}{\partial x_i}D_{ij}\frac{\partial C}{\partial x_j} + \int_\Omega wR\vartheta C \quad (A7)$$

$$= \int_\Omega wQ_C - \int_\Gamma wq_{n_C}^{total}$$

for the *divergent form* and

$$\int_\Omega wR_d\frac{\partial C}{\partial t} + \int_\Omega wq_i^f\frac{\partial C}{\partial x_i} + \int_\Omega \frac{\partial w}{\partial x_i}D_{ij}\frac{\partial C}{\partial x_j} + \int_\Omega w(R\vartheta + Q_\rho)C \quad (A8)$$

$$= \int_\Omega wQ_C - \int_\Gamma wq_{n_C}^{disp}$$

for the *convective form*, respectively, to solve the concentration C.

Weak form of the heat transport equation (17-4)

Similar to the above, the weak formulation of the convective form of the heat transport equation is given by

$$\int_\Omega w[\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s]\frac{\partial T}{\partial t} + \int_\Omega w\rho^f c^f q_i\frac{\partial T}{\partial x_i} \quad (A9)$$

$$+ \int_\Omega \frac{\partial w}{\partial x_i}\lambda_{ij}\frac{\partial T}{\partial x_j} = \int_\Omega wQ_T - \int_\Gamma wq_{n_T}^{cond}$$

for solving the temperature T.

Finite element formulations

Employing the Galerkin version of the FEM (GFEM), $w_m = N_m$, for the above weak formulations and replacing the h, q_i^f, C and T variables by their trial approximations

$$\left.\begin{aligned}
h(x_i, t) &\cong \hat{h}(x_i, t) = \sum_m N_m(x_i)h_m(t) \\
q_i^f(x_i, t) &\cong \hat{q}_i^f(x_i, t) = \sum_m N_m(x_i)q_{im}^f(t) \\
C(x_i, t) &\cong \hat{C}(x_i, t) = \sum_m N_m(x_i)C_m(t) \\
T(x_i, t) &\cong \hat{T}(x_i, t) = \sum_m N_m(x_i)T_m(t)
\end{aligned}\right\} \quad (A10)$$

the matrix coefficients of eqn (17-8) are as follows:

$$\left.\begin{aligned}
O_{mn} = \int_\Omega S_o N_m N_n \quad A_{mn} = \int_\Omega N_m N_n \quad P_{mn} = \int_\Omega R_d N_m N_n \\
U_{mn} = \int_\Omega [\varepsilon\rho^f c^f + (1-\varepsilon)\rho^s c^s]N_m N_n
\end{aligned}\right\} \quad (A11)$$

$$S_{mn} = \int_\Omega K_{ij}f_\mu \frac{\partial N_m}{\partial x_i}\frac{\partial N_n}{\partial x_j} \quad (A12)$$

$$D_{mn} = \int_{\Omega}\left[-\frac{\partial N_m}{\partial x_i}\hat{q}_i^f N_n + D_{ij}\frac{\partial N_m}{\partial x_i}\frac{\partial N_n}{\partial x_j} + R\vartheta N_m N_n\right] \quad \text{divergent form}$$

$$D_{mn} = \int_{\Omega}\left[N_m\hat{q}_i^f\frac{\partial N_n}{\partial x_i} + D_{ij}\frac{\partial N_m}{\partial x_i}\frac{\partial N_n}{\partial x_j} + (R\vartheta + Q_\rho)N_m N_n\right] \quad \text{convective form}$$

(A13)

$$L_{mn} = \int_{\Omega}\left[N_m\rho^f c^f\hat{q}_i^f\frac{\partial N_n}{\partial x_i} + \lambda_{ij}\frac{\partial N_m}{\partial x_i}\frac{\partial N_n}{\partial x_j}\right] \tag{A14}$$

$$F_m = -\int_{\Omega}\frac{\partial N_m}{\partial x_i}K_{ij}f_\mu\left[\frac{\bar{\alpha}}{(C_s - C_o)}(\hat{C} - C_o) - \bar{\beta}(\hat{T} - T_o)\right]e_j$$
$$-\int_{\Omega}N_m\hat{q}_i^f\left(\frac{\bar{\alpha}}{(C_s - C_o)}\frac{\partial\hat{C}}{\partial x_i} - \bar{\beta}\frac{\partial\hat{T}}{\partial x_i}\right)$$
$$-\int_{\Omega}N_m\varepsilon\left(\frac{\bar{\alpha}}{(C_s - C_o)}\frac{\partial\hat{C}}{\partial t} - \bar{\beta}\frac{\partial\hat{T}}{\partial t}\right) - \int_{\Gamma}N_m q_{n_h} + \int_{\Omega}N_m Q_\rho \tag{A15}$$

$$B_{im} = -\int_{\Omega}N_m K_{ij}f_\mu\frac{\partial\hat{h}}{\partial x_j} - \int_{\Omega}N_m K_{ij}f_\mu\left[\frac{\bar{\alpha}}{(C_s - C_o)}(\hat{C} - C_o) - \bar{\beta}(\hat{T} - T_o)\right]e_j \tag{A16}$$

$$\left.\begin{array}{l}R_m = \int_{\Omega}N_m Q_C - \int_{\Gamma}N_m q_{n_C}^{total} \quad \text{divergent form}\\[2mm] R_m = \int_{\Omega}N_m Q_C - \int_{\Gamma}N_m q_{n_C}^{disp} \quad \text{convective form}\end{array}\right\} \tag{A17}$$

$$W_m = \int_{\Omega}N_m Q_T - \int_{\Gamma}N_m q_{n_T}^{cond} \tag{A18}$$

Appendix B

Global smoothing of discontinuous velocity fields

A global approximation of the smoothed Darcy velocities can be written as

$$q_i^f(x_i) = \sum_m N_m(x_i)q_{im}^f \tag{B1}$$

Assume that we have an unsmoothed (discontinuous) velocity field $q_i^{f*}(x_i)$, then the smooth function which provides a best fit in the least squares sense over the domain Ω can be obtained from a minimization of the functional

$$J = \int_{\Omega}(q_i^f - q_i^{f*})^2 \Rightarrow Min \tag{B2}$$

The minimalization procedure

$$\frac{\partial J}{\partial q_{im}^f} = \int_{\Omega}2(q_i^f - q_i^{f*})\frac{\partial q_i^f}{\partial q_{im}^f} = 0 \quad for \quad m = 1, 2, ... \tag{B3}$$

or

$$\int_{\Omega}N_m(q_i^f - q_i^{f*}) = 0 \tag{B4}$$

results in a system of linear equations to solve the smoothed velocities q_i^f, viz.,

$$\sum_n M_{mn}q_{in}^f = Z_{im} \tag{B5}$$

where M_{mn} represents the mass matrix and Z_{im} is the RHS involving the unsmoothed relations. They are formed in the finite element assembling procedure as

$$M_{mn} = \int_{\Omega}N_m N_n \tag{B6}$$

and, by inserting the Darcy velocity components, as

$$Z_{im} = \int_\Omega N_m q_i^{f*} \qquad (B7)$$

$$= -\int_\Omega N_m \left[K_{ij} f_\mu \left(\frac{\partial \hat{h}}{\partial x_j} + \left[\frac{\bar{\alpha}}{(C_s - C_o)}(\hat{C} - C_o) - \bar{\beta}(\hat{T} - T_o) \right] e_j \right) \right]$$

Note, the least square approximation of global smoothing (B7) is equivalent to the Galerkin weighting procedure (A16) in Appendix A.

A cost-effective alternative appears if the mass matrix M_{mn} is lumped by an row-sum or diagonal scaling

$$M_{mn} = \begin{cases} \int_\Omega N_m (\sum N_n) & m = n \\ 0 & m \neq n \end{cases} \qquad (B8)$$

without need to solve the linear equation system (B5). Mass lumping can be shown to be equivalent to an area-weighted averaging for nodal values.

Local smoothing of discontinuous velocity fields

Unlike global smoothing, there is an efficient way to smooth velocity fields by using only individual element information. This is termed as local smoothing[45] and provides a simple nodal averaging based on the number of elements joined at a given node (element patch). Among several approaches suggested

FEFLOW employs following two-step local technique:

(Step 1) The discontinuous velocity in each element e

$$q_{ip}^{f*e} = -K_{ij} f_\mu \left(\frac{\partial \hat{h}^e}{\partial x_j} \right. \qquad (B9)$$

$$\left. + \left[\frac{\bar{\alpha}}{(C_s - C_o)}(\hat{C}^e - C_o) - \bar{\beta}(\hat{T}^e - T_o) \right] e_j \right)$$

is computed at the Gauss points p (2 x 2 (x 2) for linear and 3 x 3 (x 3) for quadratic elements) with given approximations of the hydraulic head $\hat{h}^e$, concentration $\hat{C}^e$, and temperature $\hat{T}^e$ for element e from previous solutions.

(Step 2) The values at the Gauss points are assigned to the nearest corner node $p \rightarrow m$. Each nodal contribution is summed up and, at the end, the nodal values are averaged by their number of nodal contributions n_p from the patch sharing the node m

$$q_{im}^f = \left(\sum_e^{patch} q_{im}^{f*e} \right) / n_p \qquad (B10)$$

APPENDIX C

Auxiliary problem formulation used for budget flux computation of the convective part

The budget analysis aims at the computation of the normal convective mass (or heat) fluxes

$Q_C^R = \sum_\Gamma \int (C^R q_{n_k}^R)$. Multiplying each term of the continuity equation (17-1) by concentration C we get the weak form

$$\int_\Omega wC\frac{\partial q_i^f}{\partial x_i} = \int_\Omega wC(Q_\rho + Q_{EB}) - \int_\Omega wCS_o\frac{\partial h}{\partial t} \qquad \text{(C1)}$$

It is further

$$\frac{\partial(wCq_i^f)}{\partial x_i} = Cq_i^f\frac{\partial w}{\partial x_i} + wC\frac{\partial q_i^f}{\partial x_i} + wq_i^f\frac{\partial C}{\partial x_i} \qquad \text{(C2)}$$

Employing the **divergence theorem** on the LHS of identity (C2) we obtain from (C1) and (C2)

$$\int_\Gamma wCq_i^f n_i = \int_\Omega C\frac{\partial w}{\partial x_i}q_i^f + \int_\Omega w\frac{\partial C}{\partial x_i}q_i^f \qquad \text{(C3)}$$
$$+ \int_\Omega wC(Q_\rho + Q_{EB}) - \int_\Omega wCS_o\frac{\partial h}{\partial t}$$

It has been found to evaluate the individual terms of eqn (C3) in different ways. While the velocity q_i^f in the first term of the RHS is expressed by the Darcy law, the second RHS term uses explicitly the velocity from the computation. The LHS surface integral describes already the desired convective mass flux along the boundary portion Γ, where $q_{n_k}^R = q_i^f n_i|_R$ is the normal fluid flux and $q_{n_c}^R = Cq_i^f n_i|_R = Cq_{n_k}|_R$ is the normal convective mass flux through the boundary.

Finally, following finite element formulation results to compute the normal convective mass flux from given solutions (A10) of hydraulic head $\hat{h}$, Darcy flux $\hat{q}_i^f$, concentration $\hat{C}$, and temperature $\hat{T}$

$$\int_\Gamma N_m q_{n_c}^R = -\int_\Omega K_{ij}f_\mu\frac{\partial N_m}{\partial x_i}\frac{\partial N_n}{\partial x_j}(N_k C_k)h_n$$

$$-\int_\Omega K_{ij}f_\mu\frac{\partial N_m}{\partial x_i}(N_k C_k)\left[\frac{\bar{\alpha}}{(C_s - C_o)}(\hat{C} - C_o) - \bar{\beta}(\hat{T} - T_o)\right]e_j \qquad \text{(C4)}$$

$$+ \int_\Omega N_m\left(\frac{\partial N_n}{\partial x_i}C_n\right)(N_k q_{ik}^f) + \int_\Omega N_m(N_n C_n)(Q_\rho + Q_{EB})$$

$$-\int_\Omega N_m(N_n C_n)S_o\left(N_k\frac{\partial h_k}{\partial t}\right)$$

All contributions of mass flux are summed up at node m to obtain the lumped nodal balance mass flux Q_C^R in the form

$$Q_C^R = -\int_\Gamma q_C^R = -\sum_e \int_{\Gamma^e} N_m q_{n_c}^R \qquad \text{(C5)}$$

which is defined positive inward and will be used for the boundary constraint control (see Section 17.6). Similar expressions to (C4) and (C5) can be derived for heat balance fluxes if $\rho^f c^f T$ is used as multiplier.

17. Coupled groundwater flow and transport: Thermohaline and 3D convection systems

Subject Index

Subject Index

Subject Index

Subject Index

nodes 168

O

operator splitting 226–230
orthogonal transformation 175
ORTHOMIN 78, 328
outflowing boundary 124, 141
overland flow 151, 159

P

particle tracking method 139
patch source problem 247
pathlines 141
Peclet number 237, 239, 245, 246, 268
perched water 70, 272
    problem 51
    table 39, 60
Petrov-Galerkin least square (PGLS) method 225
phase
    fluid 198, 211
    fraction of fluid 223
    gas 213
    indicator 198, 199
    interfaces 157
    lag 32
    lead error 27, 61
    multi 101, 198, 199
    solid 198, 211
    two 200
phreatic 149, 151, 164, 165, 166
    aquifer 70
    condition 154
Picard 36
    iteration 10, 22, 85, 101, 233
    modified method 26
    scheme 17
piezometric head 12, 67, 133, 134, 153, 214, 223, 266, 310
    saltwater 138
pore size distribution index 12, 46, 68, 114
pore velocity 139, 213, 226, 236, 237, 252, 265, 267

pore-connectivity parameter 12
porosity 12, 152, 199
porous media 16, 149, 151, 157, 163, 197, 199, 209, 213
predictor-corrector 21, 103, 324
    approach 11
predictor-corrector method 169
predictor-corrector one-step Newton 18–22
predictor-corrector strategy 229, 230
pressure 134, 156, 160, 213, 281, 283
    gas 214
    square 216
pressure gradient 222, 281
pressure head 12, 102, 219, 235
    air-entry 12, 27
primary variable switching 11, 15–18, 60, 101
problem
    density-driven 113
products 201
projection techniques 281
pseudo-unsaturated approach 94
pseudo-unsaturated modeling approach 83–84

R

Rayleigh number 299
    critical 329
    effective 331
    solutal 328
    thermal 328
reactants 201
reaction rates 201
representative elementary volume (REV) 152
residence time 139
residual error criterion 105
retardation 205
reverse flow field 144
Reynolds number 157
Richards equation 9, 84, 101, 233
    steady-state linearized 219
Ritz-Galerkin method 243

Subject Index

Author Index

Author Index