Hans-Jörg G. Diersch

FEFLOW

Finite Element Modeling of Flow, Mass and Heat Transport in Porous and Fractured Media

Springer

FEFLOW

Hans-Jörg G. Diersch

FEFLOW

Finite Element Modeling of Flow, Mass and Heat Transport in Porous and Fractured Media

Springer

Hans-Jörg G. Diersch
Groundwater Modelling Centre
DHI-WASY GmbH
Berlin, Germany

ISBN 978-3-642-38738-8 ISBN 978-3-642-38739-5 (eBook)
DOI 10.1007/978-3-642-38739-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013950167

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to my father, Willy.

Preface

Flow, mass, and heat transport processes in nature and geosphere are highly (if not even most) complex. There is an increasing demand in studying and predicting such kind of problems in an environmental and geohydrodynamic context. This demand naturally results from the growing human influence on the environmental and natural resources with their constraints and consequences. Men are also looking for new technologies of exploiting geothermal energy and storing fluids in reservoirs. Industries are developing new materials with improved properties for which a greater understanding of flow and energy transport is required. Among all of these applications, a very important subclass of processes occurs in structures which are categorized as *porous and fractured media*. Those structures exist in many natural and man-made systems having length scales differing by several orders of magnitude. Lengths range from pore and fracture scales in the order of micrometers and millimeters, textile and tissue materials measuring tens of millimeters, the diameter of wells in the order of tens of centimeters, the thickness of aquifer layers and geologic strata in the order of meters to tens of meters, the distances between wells and thicknesses of aquifer systems with tens to hundreds of meters, and the extent of reservoirs and subsurface fields up to tens or even hundreds of kilometers. Heterogeneities and parameter contrasts have to be encountered in all these length scales.

To understand the processes, to make them predictable and controllable, we need models. Models are abstractions of the real systems. However, abstractions are not to be considered as our resort and insufficiency in finding a description for *all* phenomena and influences. They represent a necessary and appropriate level of reduction and idealization where the (most) important processes are emphasized and the subordinate processes are dropped. This is the way (and obviously the only way) to find causal relationships and to set up predictive tools. We don't need a second perfect copy of nature; we have it already in the form of our experiments and observations made at the real system.

The construction of the model is the first and very important step in a modeling process. It is termed as model *conceptualization* which covers the description of the system's composition, the physical and physicochemical phenomena, and the relevant properties of the medium in which they occur. Obviously, such a description includes assumptions and simplifications which are subjectively selected by the modeler and, consequently, it reflects his understanding and faculty of the matter in a specific scope of interest. Accordingly, a model as a simplified version of reality is subjective, and nonunique models exist in dependence on the level of assumptions, contexts of intended applications, and the state of knowledge. Fortunately, at present model conceptualization can be based on an advanced and general framework of physics and rational thermodynamics, allowing us to objectify the modeling approach for a large range of applications. However, this requires that the modeler is conversant with these conceptual steps and understands the basic physical/thermodynamic principles of the model in order to, at least, examine the physical background of the model with its assumptions and limitations. Nowadays there is a desire to develop models (family of models) which cover a wide range of applications.

The second step in modeling is the mathematical representation of the conceptual model in the form of numerical schemes and discrete solution techniques. There are many ways to do that. However, for satisfying also the requirements of a wide range of applications as stated above for the conceptual working step, one of the best choices is the *finite element method* (FEM). The FEM is very general and useful for practical applications. Its geometric flexibility and the ability to accurately apply the appropriate boundary conditions on complex domains make the FEM superior to other numerical strategies, such as finite difference methods (FDMs) or finite volume methods (FVMs). The understanding of the actually used spatial and temporal discretization techniques is necessary for modelers who practically solve flow and transport problems and interpret the numerical results with respect to accuracy and reliability of the achieved simulation results.

The third (and final) step of modeling is the computational realization of the model (family of models) in the form of an appropriately developed *simulation software*. The graphical interface of such a simulation code represents the 'working shell' for the modeler dealing with the preparation of the input data and the execution and the evaluation of the computational results of a model. Since the software interface is the only visible and operational part of the modeling process, it can be seductive for a common or novice user to exclusively apply the software as a black box, widely ignoring the theoretical modeling basis. There is indeed a potential danger for an uncritical use of modern software. Here, a graphically very sophisticated computation can create the false impression that the quality of the numerical solution is comparable to the quality of the graphical presentation. (But the reverse of this statement is also not true: A crude graphical presentation does not necessarily indicate proper solutions.)

From the above it becomes obvious that the modeling of flow and transport processes encountered in porous and fractured media has, at least, three important faces: the conceptual, the numerical, and the software/application aspect. An "ideal" modeler should have best knowledge of all of these three subjects. But this book is not primarily addressed to such a type of a "perfect" modeler (if ever it exists), but I think, at least, both the basic concepts and the practical aspects should be reasonably well known and understood by engineers, applied scientists, and practitioners who use or intend to use models for simulating flow and transport processes in porous and fractured media.

This book is written, on the one hand, for expert modelers in this field to make the theoretical basis more understandable. On the other hand, it is also written for novices and practitioners who make contact with the matter as a software user for the first time and (hopefully) intend to improve their understanding and knowledge of the modeling basis. As the title of the book could indicate, the book is not intended as a user's guide, at least in the common sense, which would mainly emphasize software functionalities and handling. On the other hand, "real" modeling, if going into practice, should necessarily be concrete and the modeler has to decide for a specific software package (sometimes more than one). The software, which is related to this book, is FEFLOW® [125].

FEFLOW is an acronym of *finite element subsurface FLOW simulation system* and solves the governing flow, mass, and heat transport equations in porous and fractured media by a multidimensional FEM for complex geometric and parametric situations including variable fluid density, variable saturation, free surface(s), multispecies reaction kinetics, non-isothermal flow, and multidiffusive (thermohaline) effects. It is capable of handling a wide spectrum of problems ranging from theoretical studies to practical real-site applications. To master all of these supported problem classes and model options, a large degree of experience and detailed information are needed. FEFLOW comprises theoretical work, modeling experience, and simulation practice from a period of about 40 years (Table 1). In this light, the main objective of this book is to share this achieved level of modeling with all required details of the physical and numerical background with the reader. The FEFLOW book is a theoretical textbook and a reference guidance for modeling in one piece – in one hand. The theoretical basis of modeling is thoroughly described but will not stand alone; it becomes really accessible and applicable with FEFLOW. That is what I advocate and actually provide with this book: *modeling that works*.

The book is intended to put advanced theoretical and numerical methods into the hands of modeling practitioners for porous and fractured media. It starts with a more general theory for all relevant flow and transport phenomena on the basis of the continuum approach, systematically develops the basic framework for important classes of problems (e.g., multiphase/multispecies flow and transport phenomena, unsaturated-saturated problems, free-surface groundwater flows, aquifer-averaged equations), introduces finite element techniques for solving the basic 3D and 2D balance equations, in detail discusses advanced numerical algorithms for the resulting nonlinear and linear problems (e.g., adaptive techniques, variable switching strategy, upwinding schemes), and completes with a number of benchmarks, applications, and exercises to illustrate the different types of problems and ways to tackle them successfully (e.g., flow and seepage problems, unsaturated-saturated flow, advective-diffusion transport, saltwater intrusion, geothermal and thermohaline flow). All examples can be rerun, modified, and extended by using FEFLOW.

The chapters of the book can formally be grouped into two major parts: physical basis and numerical basis with benchmarks and applications. The book is not meant

Table 1 Major historical stages of FEFLOW development

Year/period	Issue
1979	*Birth* and first manifestation [128] based on the finite element predecessor program FINEL developed since 1973 [126, 127, 142]
1979–1986	*Version 1.* FORTRAN research-oriented batch program; implementations for mainframes IBM 370, EC 1055, BESM-6 with punch card input and hardcopy printed output; limited pre- and postprocessing; FEFLOW already provided an extended finite element library (quadrilaterals and triangles of linear, quadratic, or cubic type) and was able to compute 2D transient groundwater flow and transport problems [129]. Effort in modeling variable-density flow problems was initiated [130, 133]
1987–1990	*Version 2.* First interactive prototype for SUN workstations and ATARI ST microcomputer. The code was completely rewritten from FORTRAN into C. FEFLOW became the first fully interactive and graphics-based finite element simulator in groundwater [134]
1990–1992	*Version 3.* Starting commercial development. X Window System and OSF/Motif GUI implementation, installations on various UNIX graphics workstation platforms (e.g., SGI, SUN, IBM, HP, Sony, DG, DEC). Extension to 3D (1992). FEFLOW became a registered trademark (1992)
1992–2001	*Version 4.* Considerable software extension, among others: thermohaline transport modeling (1993), 3D visualization tools and GIS interfacing (1995), adaptive meshing and data store manager (1996), unsaturated flow modeling (1997), MS Windows 95/NT installation (1997), IFM programming interface (1998), and integration of parameter estimator PEST and nonlinear dispersion (2000). In the extensions of the code object-oriented programming with C++ became increasingly present
2002–2009	*Version 5.* Further advances: discrete feature elements and extended possibilities for unsaturated flow (2002); fast TRIANGLE [475] mesh generator, algebraic multigrid (SAMG) [499] equation solver (2003); multispecies transport, reaction kinetics editor, transient pathline computations, FEFLOW Explorer for 3D visualization and animation (2005); 64-bit technology, variable-density multispecies multidiffusive transport, new mesh generator GRIDBUILDER [369], scatter plots, expression editor for sink/sources (2006); and borehole heat exchanger simulation, spline interpolation, improved parallelization (2008)
2009–2012	*Version 6.* New *Qt*-based graphical interface replaced the classic X11 and OSF/Motif GUI providing a modern and powerful environment for modeling and simulation available for both MS Windows and LINUX operation systems. GUI, data management, and part of the computational finite element kernel were transformed to a rigorous object-oriented architecture based on C++
2012–...	*Version 6.1.* Completion of the new object-oriented software architecture with *Qt*-based GUI. 3D sterioscopic graphics available. Improvements in parallel computing and high performance in large data treatment and simulation

to be read from front to back. The first part can also be of interest for those readers who wish to learn more about continuum mechanics for flow and transport phenomena in porous and fractured media. Others could primarily be interested in the finite element method with the embodied numerical algorithms. However, I assume most readers will start up with a software play and will hopefully be more interested in the basics later on (as the inductive way of learning – “from the surface into the ground”). To support this approach, I endeavor to present the subject in a complete and unified manner. At the beginning of the book, the preliminary chapter will summarize all important notations, definitions, and fundamental algebra used throughout the text.

I hope the book will be useful for both students and practitioners in engineering and geosciences as well as in other fields where porous-media flow dynamics and computational methods are of specific concern. I suppose that the reader already possesses (or approaches) an advanced degree in engineering or applied sciences and has an interest in geohydrodynamic flow modeling. I assume that the reader is somewhat versed in physical/mechanical principles and numerical mathematics.

Berlin, Germany
March 2013

Hans-Jörg G. Diersch

Acknowledgments

This book accumulates modeling work over a period which I started in 1973 at the Dresden University of Technology, Germany, with finite elements. In 1979 I began to focus on groundwater and porous-media flow modeling and started the FEFLOW development at the former Institute of Mechanics in Berlin (Academy of Sciences of East Germany). I spent 12 fruitful years at this Academic Institute in computational fluid mechanics, which I deeply acknowledge. In 1990 I founded the WASY company together with partners (in 2007 renamed to DHI-WASY GmbH) in Berlin, Germany, where FEFLOW has been further developed, continuously improved, and extended. Since that time the capabilities of the code have been increased substantially. I thank the FEFLOW developing group at DHI-WASY for their enthusiasm and contributions. With the book I would like to express my gratitude to all previous and current colleagues who are directly or indirectly involved in achieving the success of FEFLOW as a scientific and commercial software. I hope the book can provide a secure and fertile basis which I would like to hand over to the next generation of FEFLOW developers.

My most special appreciation is dedicated to my wife Renate who has accompanied my work since the time when FEFLOW consisted of 500 punched paper cards, ran on a computer bigger than a house, and the only available plotter was my own hand (drawing velocity arrows of about 1,000 nodes!). She is the only one who can really mark the miles I went through the times.

I am grateful to Fabien Cornaton and Julia Mayer (DHI-WASY) for reading the manuscript and offering constructive suggestions. My thanks are also due to Christian Witschel, Editorial Director Earth Science of Springer Verlag, and Venkatachalam Anand, Project Manager of SPi Content Solutions India, for their management of the publication process of this book.

Contents

Part I Fundamentals

Part II Finite Element Method

Acronyms

1D	one-dimensional
1U	single U-shape pipe
2D	two-dimensional
2U	double U-shape pipe
3D	three-dimensional
4D	four-dimensional
AB	Adams-Bashforth
AB/TR	Adams-Bashforth/trapezoid rule
ADE	advection-dispersion equation
AFT	advancing front technique
AMG	algebraic multigrid
AMR	adaptive mesh refinement
AREV	aquifer REV
BASD	best adaptation to stratigraphic data
BC	boundary condition
BCC	boundary condition constraint
BE	backward Euler
BFGS	Broyden-Fletcher-Goldfarb-Shannon
BHE	borehole heat exchanger
BiCGSTAB	bi-conjugate gradient stabilized
BTEX	benzene-toluene-xylene mixture
CAD	computer-aided design
CBFM	consistent boundary flux method
CFD	computational fluid dynamics
CFL	Courant-Friedrichs-Lewy
CG	conjugate gradient
CGS	conjugate gradient square
CM	consistent mass
CR	constitutive relation
CSA	cold and salty above
CSB	cold and salty below

CXA	coaxial pipe with annular inlet
CXC	coaxial pipe with centered inlet
DBF	Darcy-Brinkman-Forchheimer
DDC	double-diffusive convection
DDFC	double-diffusive finger convection
DF	Darcy-Forchheimer
DFE	discrete feature element
DOF	degrees of freedom
DVM	Delaunay-Voronoï method
EFG	element free Galerkin
EOB	extended Oberbeck-Boussinesq
EOS	equations of state
FDM	finite difference method
FD3DM	fully discretized 3D model
FE	forward Euler
FEFLOW	finite element flow simulator
FEM	finite element method
FKA	Frolkovič-Knabner algorithm
FPM	finite point method
FU	full upwinding
FVM	finite volume method
GFEM	Galerkin finite element method
GIS	geographic information system
GLS	Gresho-Lee-Sani
GMG	geometric multigrid
GMRES	generalized minimal residual
GUI	graphical user interface
GWS	Galerkin weak statement
HC	high concentration
HOT	higher order terms
HRL	Horton-Rogers-Lapwood
HSA	hot and salty above
HSB	hot and salty below
IC	initial condition
ID	identifier
ILU	incomplete LU decomposition (factorization)
LBB	Ladyshenkaya-Babuška-Brezzi
LHS	left-hand side
LM	lumped mass
LMA	law of mass action
LS	least square
LSGFEM	least-square Galerkin finite element method
LSWS	least-square weak statement
LTE	local truncation error
MG	multigrid

MILU	modified incomplete LU decomposition (factorization)
MLM	meshless method
MLNDS	multilevel nested dissection
MOC	method of characteristics
MPFA	multi-point flux approximation
MWR	method of weighted residuals
NURBS	nonuniform rational B-splines
OB	Oberbeck-Boussinesq
OBC	outflow boundary condition
ODE	ordinary differential equation
ORTHOMIN	orthogonalization-minimization
PDE	partial differential equation
PGFEM	Petrov-Galerkin finite element method
PGLS	Petrov-Galerkin least square
PVST	primary variable switching technique
RCM	reverse Cuthill-McKee
REV	representative elementary volume
RHS	right-hand side
RMS	root mean square
SC	shock capturing
SEM	spectral element method
SIA	sequential iterative approach
SMP	symmetric multiprocessing
SPC	singular point condition
SPCC	singular point condition constraint
SPR	superconvergent patch recovery
SSA	steady-state approximation
SU	streamline upwind
TDS	total dissolved solids
TPFA	two-point flux approximation
TR	trapezoid rule
TRCM	thermal resistance and capacity model
TRT	thermal response test
VG	van Genuchten
VGM	van Genuchten-Mualem
WS	weak statement
XFEM	extended finite element method

Chapter 1
Introduction

1.1 Scope of Modeling

Flow, mass and heat transport through *porous and fractured media* occurs in many branches of engineering and science. Of particular concern are those processes in the *subsurface* occurring beneath the surface of the earth's ground, that means flow and transport in geologic media with their complexity and uncertainty. Within the pore voids, fractures, channels, cavities and other non-solid spaces of the geologic formations the movement of fluids, solutes (chemical species) and heat is of central interest. Fluids represent a general term encompassing liquids, gases and gas-liquid mixtures. Among the different types of fluids to be encountered in the environmental and hydrogeologic context, *water* is the most important fluid. The earth's water resources, especially the available freshwater, represent the basis for human, animal and plant life and its growth of importance results from the continually increasing demand for drinking water, effects by pollution, danger by over-exploitations and climate changes.

Water on, under and above the earth's surface form the hydrologic cycle consisting of the subdomains atmosphere, sea, surface water and *subsurface water*. Only a minor part of the water volume represents freshwater (2.5 % of all water on earth). Compared with the freshwater volume stored in ice caps, lakes and rivers, the subsurface water comprises 99 % of the earth's available freshwater [119, 356]. Subsurface water is often subdivided into soil moisture of the *unsaturated zone* and *groundwater* of the saturated zone of the underground (definitions of terms are summarized in Sect. 2.2). This division is appropriate to differ between the physical processes governing the unsaturated and saturated zones. Accordingly, subsurface modeling has to consider variably saturated conditions for both flow and transport processes in the underground.

Flow and transport processes in porous and fractured media include diverse phenomena such as the spreading of toxic waste products (miscible and immiscible contaminants), movement of natural chemical constituents (e.g., saltwater), deposits of fluids and hazardous wastes, energy storage and recovery encountered in various

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_1,

environmental and industrial applications. The modeling of those flow and transport processes gives, e.g., the opportunity to:

- Describe the spatial and temporal distribution of contaminants and/or thermal fields,
- Analyze moisture dynamics and seepage processes,
- Assess irrigation and drainage potentials as well as salinization of soils,
- Study freezing and thawing processes in soils,
- Estimate the duration and travel times of fluids and pollutants,
- Plan and design remediation strategies and containment technologies,
- Assist in designing effective monitoring schemes,
- Predict groundwater-management measures,
- Predict flow and stress formations underneath engineering structures,
- Quantify flow of oil, water and gas in reservoir engineering,
- Plan a nuclear waste repository in geologic formations,
- Design and quantify drainage and flooding processes in mines,
- Design geothermal energy extraction and storage systems,
- Assess saltwater encroachment in coastal regions and saltwater upconing below pumping wells,
- Predict brine and thermohaline influences in deep locations, and
- Describe drying or absorption processes in deforming porous materials.

In the field of water management, hydrogeology, geophysics and mining industry the flow and transport processes of groundwater (i.e., in the saturated zone) clearly dominate. Their modeling is commonly based on a *single-phase approach*, where water is the only dynamic phase in which one or more chemical species are entirely dissolved. The groundwater body is often bounded at the surface by a water table and *free-surface flow* is typical. Flow and transport processes in the unsaturated zone often occur in hydrological and agricultural applications. Here, the voids are only partially filled with water, the remaining part contains gases mainly in form of air and water vapor. Basically, unsaturated flow needs two fluid phases, where dissolved components can additionally occur in both dynamic phases. However, it can often be assumed that the gas phase remains stagnant at a constant (atmospheric) pressure level and a reduction to a one-phase model is common. On the other hand, there are organic substances which possess hydrophobic properties, i.e., they are immiscible with water and only slightly soluble (e.g., petroleum products and halogenated hydrocarbons). If those contaminants intrude into an unsaturated zone three dynamic phases simultaneously occur consisting of water, air and organics. They represent a highly complex dynamical system which is often difficult to solve. Additional effort and difficulties naturally occur if the one-, two- or three-phase system with its chemical constituents is affected by thermal processes. Non-isothermal multispecies multiphase processes are the most general and indeed the most difficult problems to be encountered in subsurface flow and transport modeling.

A further typical feature of all flow and transport processes in porous and fractured media is that they occur on extremely different spatial scales ranging from a regional scale of geologic structures in orders of kilometers to the pore scale in orders of micrometers. Here, one fundamental question is how much detail should be included in the conceptualization of the basic models. This matter has received large attention in recent years. The most promising developments in this field are based on advanced *thermodynamic theories* in which a unified and general framework for constructing flow and transport models in porous and fractured media has been attained.

1.2 Interdisciplinary Aspects

The construction of flow and transport models with their necessary practical implication and simplification, their mathematical-numerical representation and coding, their accompanying data acquirement and parameter estimation, their software development and simulation, their verification and problem adaptation as well as their application and evaluation in a practical context is inherently interdisciplinary in nature which joins, among others, aspects from rational thermodynamics,[1] fluid mechanics, mathematics, chemistry, geology, geophysics, computer science and engineering. An interdisciplinary team is capable of addressing the multifaceted aspects occurring in practical modeling.

1.2.1 Continuum-Mechanical View

The continuum approach is the standard and the most successful way to describe the fundamental processes of flow, mass and heat transport in porous and fractured media. Fluid mechanics, solid mechanics and thermodynamics are subclasses of physics in which the world is viewed as a *continuum*. The assumption of a continuum means that physical properties (such as velocity, stress, temperature etc.) distribute through space and are connected with a *material point*. At the material point the properties have finite values. The properties may change from one point to the next, and there may even be surfaces where properties can jump discontinuously. However, a continuum approach does not allow properties to become infinite or to jump discontinuously at a single isolated point. As a result, such an approach introduces an effectively *continuous medium* which is characterized by a relatively small number of bulk properties, such as density ρ, compressibility γ, viscosity μ, concentration C_k of chemical species k and temperature T. Under such conditions,

[1] In thermodynamics, *rational thermodynamics* is a very general phenomenological and macroscopic theory for deriving constitutive equations, basically established by C. Truesdell [520] and his students W. Noll and B. Coleman in the 1960s, and is distinct from other categories of thermodynamics such as the classical thermodynamics and (extended) irreversible thermodynamics.

fundamental laws are written for this kind of continuous media in form of continuum balance statements which take the form of partial differential equations in space coordinates $\boldsymbol{x}$ and in time t.

A continuum represents a *macroscopic* view on physical events. This is different to events occurring in the microscopic world of molecules, nuclei and elementary particles, which are not governed by continuum laws. Continuum properties may be interpreted as the *average of events* involving a great number of microscopic particles. The transition from the microscopic to the continuum level is the subject of statistical thermodynamics (statistical mechanics) and kinetic theory. The microscopic processes can be described, at least, on two levels. At a molecular level the motion is reversible (called Hamiltonian motion), while at the kinetic level irreversible motion is formulated by the low-density *Boltzmann approximation*. Such Boltzmann models are fundamentally probabilistic, discrete in velocity, however continuous in space and time. Microscopic models based on such a statistical framework can be helpful for a deeper understanding of continuum properties and relations between microdynamics and macrodynamics, but, on the one hand, usually there is no real need for such models and, on the other hand, essential constraints exist for their use in a general and more practical context. In fact, some interesting applications of Boltzmann-type model can be found in modeling rarefied gas-flow problems by using *lattice-gas methods* (Doolen et al. [150]) and *cellular automata* (e.g., [516,567]), where collision rules for an appropriate number of particles are discretely simulated without the necessity for standard numerical techniques for solving partial differential equations. A natural transition from molecular dynamics to a continuum approach is provided by averaging techniques applied to a sufficiently large number of molecules (ensemble of particles) [488].

Under the aspect of a continuum approach flow, mass and heat transport processes in porous and fractured media appear as a subclass of fluid mechanics, extended by parts of solid mechanics, thermodynamics and chemistry as shown by the Venn diagram of Fig. 1.1. Typically in classic fluid dynamics there is a strict subdivision into incompressible and compressible fluids which has some mathematical consequences. While a fluid is said to be incompressible if the density of a fluid continuum is not affected by pressure changes, the term ‘incompressible flow’ means that density variations in the conservation of fluid mass (continuity equation) are neglected. We shall see later on that flow and transport problems in porous and fractured media have to consider the compressibility conditions of the whole multiphase system. Therefore, flow and transport in porous and fractured media generally describes *compressible* phenomena.

1.2.2 Mathematical View

From the mathematical point of view the main objective is the solution of the governing continuum balance equations for given initial and boundary conditions. The subject of the mathematical analysis is termed as the *mathematical model*

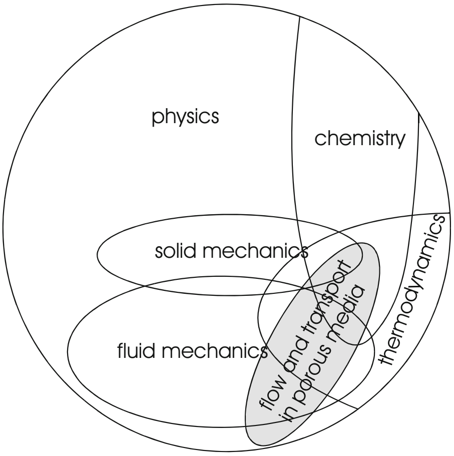


Fig. 1.1 Flow and transport in porous media overlaps different scientific disciplines: an attempt of classification

which is described by a set of partial differential (or integro-differential) equations. However, for the most problems of interest exact or analytical solutions of the basic mathematical model are neither available nor attainable and, accordingly, *approximate* methods are commonly required.

It is to be emphasized that a mathematical model already contains idealizations and approximations of 'reality'. This is a consequence of assumptions and simplifications necessarily being made in the derivation of the basic continuum equations and constitutive relationships (the *conceptual model*) to make the model 'tractable'. Now, the mathematical solution itself requires approximations to be made by the discretization process. Additionally, the solution of the discretized equations can introduce further errors, for instance if solving the resulting nonlinear or linear equation systems by iterative methods. Different sources of *errors* can arise:

(1) Errors embedded in the conceptual model resulting from assumptions and simplifications in deriving the basic mathematical model equations. Even if it is possible to solve these equations exactly, the solution would not be a correct representation of reality. To estimate the *modeling errors* a *validation* of the model is required, where comparisons with experimental and observation data have to be performed. The validation process is a topic in itself. We should bear in mind that detected errors are not necessarily caused by applying improper model equations. Often, reasons have to be searched in the model parameters (the constitutive dataset) and/or in inappropriate boundary conditions. Furthermore, we should not forget that measurements can possess their own errors.

(2) *Discretization errors* result from the use of numerical methods to solve the basic model equations. We should select techniques which allow to keep the discretization errors small. Surely, it would be best to reduce the error as much as possible. But, this is not practical and, in fact, not always necessary. More accurate approximations can dramatically increase the time and cost of obtaining the solutions. Compromises are usually needed. This is a question of

optimality: A refined discretization (and a higher computational effort) is only necessary in such regions and/or at such times where (in which) mathematically required. So we look for solutions possessing an optimal accuracy.

(3) The solution of the discretized equations needs further compromises. The nonlinear equations have to be linearized by *iterative techniques*. Furthermore, iterative techniques have to be applied in solving the resulting (often large) sparse matrix equation systems because exact (direct) solvers are too costly and only applicable to smaller problems. In all, the iterative procedures must be carefully controlled and terminated after satisfying prescribed convergence criteria as a measure of tolerated errors (*convergence errors*).

It is obvious that a final model realization (called simulation) is affected by all of those errors. From a mathematical point of view errors stemming from model conceptualization and parameter input are very complex and normally out of the mathematical scope. Now, we can argue: If we are capable of measuring (and controlling) the discretization errors at an optimally small level we can exclude (better we minimize) the influences of numerical errors in the model simulations. If we then compare our model results to experiments and observations we can be sure that occurring differences are no more caused by an improper discretization. Indeed, this is a high goal in numerical mathematics in developing reliable and robust schemes. Such numerical schemes should have certain properties, e.g., [162]:

(i) CONSISTENCY. The discretization should become exact as the mesh spacing $\Delta \boldsymbol{x}$ and temporal increment Δt tend to zero. In calling the difference between the discretized equations and the exact solution as the *truncation error*, a method is *consistent* if $\Delta \boldsymbol{x} \rightarrow \boldsymbol{0}$ and/or $\Delta t \rightarrow 0$.

(ii) STABILITY. A solution method is said to be stable if it does not magnify errors during a numerical solution process. For transient problems, stability guarantees that the method produces a bounded solution whenever the solution of the exact equation is bounded. For iterative methods, a stable method is one that does not diverge. It is to be remarked that stability does not imply accuracy – although it is true that instability implies inaccuracy.

(iii) CONVERGENCE. For a scheme which satisfies the consistency condition we additionally require stability as a necessary and sufficient condition for convergence.

(iv) CONSERVATIVITY. Since the model equations represent balance equations for conserving physical quantities, the numerical scheme should accurately satisfy this statement of balance on both a local and a global basis. Conservativity is a very important and most fundamental property required for the proposed numerical schemes.

(v) BOUNDEDNESS. Numerical solutions should lie within proper bounds. Physically non-negative quantities (e.g., density, concentration, absolute temperature) must always be positive. Boundedness is difficult to guarantee. Unbounded solutions can occur on too coarse meshes or too large time steps in form of *wiggles* exhibiting overshoots and undershoots of the solution. Wiggles are usually a signal that the spatial discretization is too coarse and some refinements (at least locally) are required. Stability and convergence problems can result if the solutions are too much wiggled. However, a positive aspect of wiggles is that they signal improper discretization and, accordingly, possess a *self-diagnosis property* [209]. A method with such a self-diagnostic property is often superior to schemes which give smooth and totally wiggle-free, but inaccurate and overdamped solutions for each discretization. We find, boundedness and accuracy are connected in a strong and often contrary manner.

(vi) ACCURACY. A scheme works accurately if both the discretization and the convergence errors remain sufficiently small. Unfortunately, in many practical applications modeling (conceptual) errors are an additional concern. The separation of modeling errors from 'true' numerical errors appears as an important task because various errors may cancel each other, so that sometimes a solution obtained on a coarse mesh may agree much better with the experiment than a solution on a finer mesh – which, by definition, should be more accurate. It is obvious that models and algorithms have to be tested under different aspects to quantify the order of accuracy. Such tests are categorized as follows [20]:

- *Verification:* A comparison to a problem which is sufficiently elementary such that the analytical solution is known.
- *Benchmark:* A comparison to a problem which possesses the intrinsic physical and mathematical character (e.g., nonlinearity) of the basic model, but applied in a simplified geometry such that comparative numerical solutions of accepted (known) quality are available.
- *Validation:* A comparison to a problem for which quality experimental data are available.

There are many numerical methods for approaching the basic modeling equations. The most important strategies are: the finite difference method (FDM), the finite volume method (FVM) and the finite element method (FEM). Other methods such as spectral schemes, boundary element methods, global meshless techniques and cellular automata are limited to special classes of problems. Among the above methods the FEM and the FVM are the most powerful methods. While the FDM approximates the differential form of the basic balance equations in a difference form and is restricted to simple geometries and boundary conditions, both FEM and FVM are based on the weak, variational formulation of the boundary and initial value problem, where the solution appears in the integral of a quantity over an arbitrary domain. This integral approach – in contrast to the difference (differential) approach of FDM – is the actual power of FEM and FVM, which is a natural and an adequate approach of a continuum balance statement. Indeed, the balance laws of continuum mechanics are global in the sense that they are integral laws applied to a given mass of material, fluid or solid. FEM and FVM subdivide the continuum in a finite number of elements (FVM says control volumes), for which the balance statements are discretely applied. There are more similarities than differences between FEM and FVM, however, the FEM appears to be the

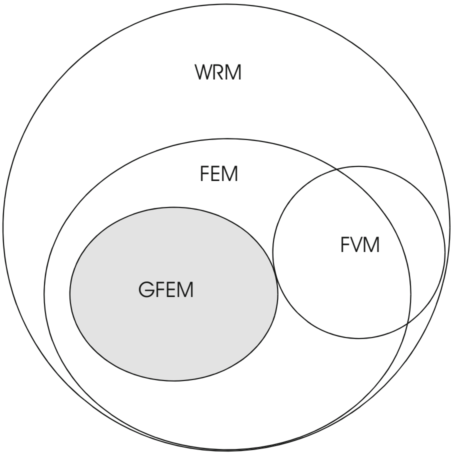


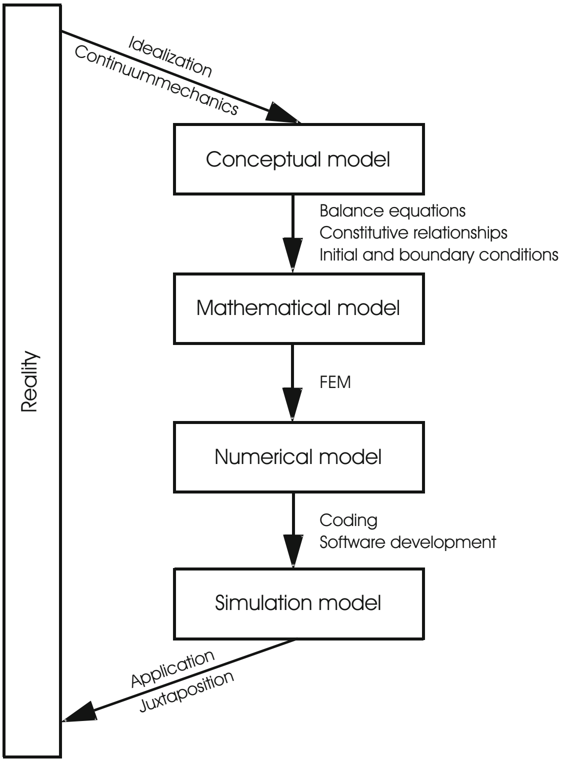
Fig. 1.2 The FEM/GFEM and FVM are submethods of the WRM

most general and powerful method. FEM is superior to the others due to following features (see, e.g., [84, 209]):

(a) *Arbitrary geometries.* The FEM is essentially geometry-free. In principle, FEM can be applied to domains of arbitrary shape and with quite arbitrary boundary conditions.
(b) *Unstructured meshes.* FEM by its nature leads to unstructured meshes. This means, in principle, modelers can place finite elements anywhere they please. Accordingly, most complex types of geometries can be simply handled.
(c) *Robustness.* In the FEM the contributions of local approximations over individual elements are assembled together in a systematic way to achieve a global approximation of a solution to a partial differential equation. Generally, this leads to schemes which are stable in appropriate norms and, moreover, insensitive to singularities or distortions of the mesh, in sharp contrast to the classic FDM.
(d) *Mathematical foundation.* Today, a solid and rich mathematical basis is available for the FEM. It covers methods to determine a priori and a posteriori error estimates and helps to advance the FEM for important (and new) application problems above a traditional level of empiricism.

It has been demonstrated by Gresho and Sani [209], additionally by [83, 284], that the FVM is inherently a FEM if using low-order elements (basically linear). O.C. Zienkiewicz (noted in [209]) stated: *The FVM is a poor-man's FEM; it's a FDM moved over half-way*. To the end, the (Galerkin-based) FEM (GFEM) appears as a generalized FVM. The FVM is (often if not always) also a weighted residual method (WRM) [163]; only the weighting functions are different (Fig. 1.2). For more discussions, see Chap. 8.

Fig. 1.3 Stages of model development



1.2.3 Computer-Scientific View

With the help of continuum mechanics and rational thermodynamics we construct a *conceptual model* and end up with a mathematical model possessing balance statements for physical quantities and a set of constitutive relations (Fig. 1.3). The *mathematical model* becomes solvable after transferring it into an appropriate *numerical model*, e.g., by employing FEM. To perform a numerical model for practical needs it has to be run on a computer. For this purpose the numerical model is appropriately coded by using programming techniques (computerization). At the end, a simulation program (sometimes said *simulation model*, *simulator* or *simulation system*) results which allows the solution of the basic balance equations for different problem types, geometries, time ranges, parameter situations, initial and boundary conditions. In addition, we should also be able to run it for different approximation levels in varying spatial and temporal resolutions as well as the types and alternative strategies of numerical schemes embodied in the simulation code.

At first glance, it seems to be sufficient to code (implement) strictly the numerical model with its encountered variants of algorithms, procedures and solution strategies. Indeed, this is fundamental but we have to ask ourselves whether this is enough

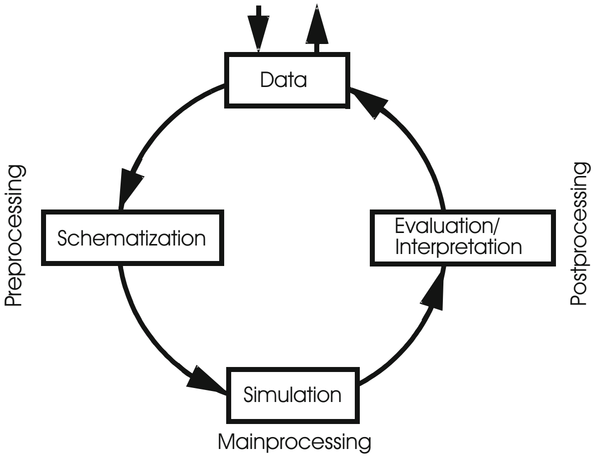


Fig. 1.4 Modeler's working loop

for practical requirements. In designing a simulation software we have to answer the following questions:

- For whom should it be developed? It makes a difference whether the code is primarily designed for teaching, purely scientific and research-oriented or daily-life practical needs.
- Which range of problems and applications should be covered? We are realistic and believe that is not possible to meet *all* needs in one *piece* of software.
- Which sources of data and information needed for the model should be interfaced? This refers to the question of confidence, efficiency, completeness, detailedness, reliability and repeatability in the practical simulation work using real data; in a nutshell *simulation fidelity*.
- How should the access to and exchange of input data and computational results be designed? A simulator should feature interactive graphics, sophisticated visualization, multiprocessing and parallel computation to tackle the simulation challenges at present and future times.

It becomes clear that a useful simulator for flow and transport processes in porous and fractured media has to be equipped with a number of additional features beyond the pure analysis core of the numerical model. Usually, a modeler is continually faced with the four major working steps as schematized in Fig. 1.4. Data of different kinds and from different sources have to be collected and analyzed. They can be both primary and secondary information of a real problem to be studied (point samples, profiles, maps, databases, 3D geometries, comparative scenario data, etc.). Based on these data the modeler builds up a schematization of the real conditions, where appropriate geometric and parametric idealizations are performed in the objective of the intended modeling. This preprocessing step covers two different aspects of work. The first one refers to the ingenuity, ability and creativity in abstracting the real-world conditions: What is important? What can be dropped? Where are the right borders for the study domain? and so on. The secondary aspect is fully technical/technological: Are there tools for an efficient meshing and parameter assignment (parametrization, regionalization)? May the code handle model data quite independently of a discretization? Is the editing process interactive and does it allow an easy manipulation of all important model data? and others.

Next, the modeler performs the simulation under the specified conditions and parameters. If the simulation finishes successfully, the computational results have to be evaluated and interpreted in different ways. Graphics and visualization tools support the modeler in this important phase of modeling to tackle the often voluminous modeling output. The achieved computational results are to be compared to and interfered with the basic and measurement data (*juxtaposition process*). As a result, this normally feeds back to a repeated computation by restarting the design and analysis loop.

The designing, simulation and evaluation/juxtaposition processes require efficient and powerful tools in data handling, manipulation, computation and visualization. There are two major reasons for their emphasizing. First, the pre- and postprocessing work is commonly the most time-consuming and error-sensitive task. Indeed, besides the errors arising from the conceptual and numerical approaches as stated above, there is a danger for introducing additional errors which result from mistakes in data handling and misinterpretations of the results. Simulation software should incorporate numerical and visual capabilities for supporting the detection of such errors caused by data mismatching.

Second, in practice the working loop (Fig. 1.4) has to be cycled for subsurface flow and transport processes very often due to, among others, the following reasons:

- Uncertainties in the database (e.g., spatial variability of geologic information).
- Scaling effects (macroscopic parameters can be scale-dependent).
- The (customer's) need for repeatability and cross-checking of a model prediction (increasing objectivity and transparency).
- The need for scenario analysis in order to, for example
  - Detect causal dependencies,
  - Perform parameter sensibilities and model calibration,
  - Estimate field parameter and probabilistic characteristics,
  - Enforce 'epignosis' (verification, history matching),
  - Enforce prognosis under altered assumptions,
  - Assess remedial schemes, technological strategies and alternative design concepts (design computations, reverse engineering),
  - Control and design monitoring schemes, and
  - Optimize in-situ measurement programs.

On the one hand, the easier and more effective the cycle can be passed through, the more completely the problem can be studied and assessed. On the other hand, it is obvious that the simulation of flow and transport processes in porous and fractured media is not simply a straight-forward solution of the discretized balance equations

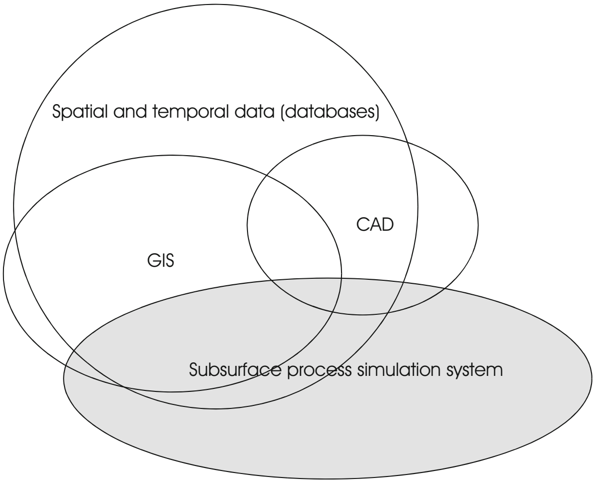

Spatial and temporal data (databases)
CAD
GIS
Subsurface process simulation system

Fig. 1.5 Venn diagram indicating that a subsurface process simulator is interfaced to GIS (or/and CAD) to give access to databases

with some parameters, initial and boundary conditions, rather more it is a *play with data*, where the simulator introduces predictive capabilities on an advanced physical basis. This requires an effective framework for visual and quantitative communication. It is supported by sophisticated Geographic Information Systems (GIS) or/and Computer-Aided Design (CAD) systems (Fig. 1.5). For managing environmental and geologic information GIS is more popular and appropriate [419], while CAD has prevailed in industrial branches and structural engineering. A simulator for subsurface processes can greatly benefit from the use of GIS (or CAD). Very important aspects of the quality of final computations are accuracy, time, scale and completeness of datasets. The effective access to data and the associated quality of data will affect the accuracy of the modeling and therefore defines the usefulness of the developed simulation system.

GIS is a tool for storing, manipulating, analyzing and displaying spatial or geographically referenced data. GIS can primarily be seen as a database. It can store, maintain, recover and update spatial data and associated descriptive information. GIS data are stored in either vector or raster forms. Vector data are sets of points, lines and polygons, while raster data are stored in a matrix of columns and rows (grid format). These data representations are sufficient for 2D applications. In contrast, 3D GIS uses volumetric data structures which are stored either by 3D boundary representations for vector data or by volume elements (voxels) for raster data.

In principle, a CAD system provides the same features as GIS. The most significant difference to CAD and other databases is the spatial nature of data in a GIS. In addition to the pure database functionality GIS provides analysis functions which allow manipulation of multiple themes of spatial data to perform overlays, buffering and arithmetic operations on the data.

Databases of geologic formations are usually affected by uncertainty and randomness. Their lithologic, petrophysical and structural features can exhibit wide variations on different scales which cannot be described deterministically in all relevant details. By using geostatistical tools [79] (such as *kriging* [357]) the parameter values typically only known at a small number of sampling (or monitoring) points are interpolated in a random field defined over the entire domain (data regionalization). Those resulting parameter distributions are probabilistic in nature and provide useful estimates in subsurface flow and transport modeling. The randomness in the response of a flow and/or transport process can further be studied via sensitivity analysis or Monte Carlo simulations or directly by solving stochastic differential equations, e.g., [38, 109, 191].

From the above, the requirements to a modern simulation system for computing flow and transport processes in porous and fractured media are rather manifold. The architecture and the programming of the simulator have to meet, among others, the following criteria:

- The simulator has a sophisticated and graphical user interface.
- The simulator has an open data interface and modular architecture.
- The code guaranties a high portability and expandability.
- The code fully dynamically manages the physical memory demand.
- The code works efficiently and fast, it allows the parallel computation of large and complex problems.
- The used programming languages provide a readable, verifiable, extensible and reusable coding.
- The program supports distributed computing on local- and wide-area networks.

The code development, maintenance and support of a simulation software, such as FEFLOW [125], is a multidisciplinary teamwork for itself. *Object-oriented programming* (like C++) and Computer-Assisted Software Engineering (CASE) tools facilitate the development and management of an extent of millions of code statements. Extensive and continuous testing of the simulation program is required to fix software errors (*bugs*) and to ensure that it is sufficiently efficient in terms of speed, required resources and usability.

It is a natural consequence that software products of this kind are developed on a commercial basis. The proliferation of powerful computers makes the computation of larger and larger problems by an increased number of users possible. The commercialization submits a distribution of software to rules of the market: *as much as possible*. It can entail unwanted side effects. Some users treat the simulation code as a black box. Results often appear plausible for nonspecialists, even when the results are grossly inaccurate. Other users tend to prefer (exclusively) the most complex solution provided by the simulator (e.g., large 3D), where causal dependencies and essentials often remain hidden (*cannot see the wood in too many trees*). The computational requirements can greatly vary from application to application (e.g., classic groundwater flow versus reactive and multiphase transport simulation). There is an allurement for software developers to attempt to satisfy all demands in one product. As a result, the software tends to become

- Highly complex,
- Difficult to use properly, and

- Slow in responding to new requirements.

A substantial amount of training is required. In fact, some users lack the required expertise. It is one of the major objectives of this book to emphasize the *trinity* of theory, technology and practice. Success and confidence in modeling of flow, mass and heat transport processes in porous and fractured media depend on the reliability of the model formulation, the reliability of the parameters used, the reliability of the numerical solution and the proper interpretation of the results.

1.3 Taxonomy for Porous/Fractured Media Process and Numerical Modeling

As stated above the modeling of flow, mass and heat transport processes in porous and fractured media by using numerical approaches is a multidisciplinary task. Solutions strategies and methods have been developed over the last 50 years which form the background and the 'state of the art' of today's modeling. A taxonomy for modeling approaches and numerical techniques can be developed to classify the developments and historical basis. The next Tables 1.1–1.4 cite and summarize the most important works and books in the field of our interest.

1.4 Overview of This Book

Part I of the book covers the fundamentals of modeling flow, mass and heat transport processes in porous and fractured media. It starts with the preliminaries in Chap. 2, where all basic definitions, expressions and principles are introduced, which will be important through the book. It can be considered as a condensed source of information on basic physical and mathematical concepts and foundations. Notations and quantities useful in subsurface and porous/fractured-media modeling are listed and described together with their theoretical and practical context. Fundamental multiphase and multispecies concepts are introduced in Chap. 3. It describes the spatial averaging method to transform microscopic quantities to macroscopic (porous-fractured medium) quantities based on the REV concept. Their balance equations for mass, thermal energy and entropy are derived in detail. Useful thermodynamic principles are described. They form the basis for developing phenomenological laws, equations of state and appropriate constitutive relationships needed. The closed set of the basic model equations is systematically derived and summarized in their different levels of reduction. Chapter 4 represents discrete-feature modeling basics. Fundamental equations are developed for diffusion-type flow (Darcy, Hagen-Poiseuille, overland), mass and heat transport. Chapter 5 is devoted to chemical reactions and kinetics. Adsorption relations for Langmuir, Henry and Freundlich isotherms and kinetic formulations of degradation, Arrhenius,

Table 1.1 Conceptual model development (fundamental theoretical work)

Author(s)	Title (year) [reference]	Comments
A.C. Eringen and J.D. Ingram	A continuum theory of chemically reacting media – I (1965) [158] and II (1967) [285]	Unified theory for derivation of balance and constitutive equations of chemically reacting (non-porous) media
S. Whitaker	The method of volume averaging (1999) [563]	'State of the art' of spatial averaging applied to single and two phase flow systems in porous media with emphasis on chemical reaction, heat transport, dispersion and heterogeneity
J. Bear	Dynamics of fluids in porous media (1972) [33];	The 'standard' book for modeling of porous media
	Modeling flow and contaminant transport in fractured rocks (1993) [35]	Derivating conceptual model equations of fractured rocks regarding flow and mass transport
J. Bear and Y. Bachmat	Introduction to modeling of transport phenomena (1991) [37]	Theoretical book of porous media giving a rigorous derivation for the most important equations of single and multiphase flow, mass and heat transport
W.G. Gray	A derivation of the equations for multiphase transport (1975) [201];	Introducing the modern methodology of spatial averaging for multiphase systems
	Derivation of vertically averaged equations describing multiphase flow in porous media (1982) [202];	Rigorous derivation of aquifer-type model equations
	Thermodynamics and constitutive theory for multiphase porous-media flow considering internal geometric constraints (1999) [203]	Providing advanced thermodynamic approach and constitutive theory for porous-media flow
S.M. Hassanizadeh and W.G. Gray	General conservation equations for multiphase systems (1979, 1980) [226–228];	Founding the general thermodynamic multiphase approach in porous media
	Mechanics and thermodynamics of multiphase flow in porous media including interface boundaries (1990) [230]	Interfacial transport phenomena
G. Dagan	Stochastic modeling of flow and transport (1997) [109]	Stochastic modeling of subsurface flow and transport problems
P.C. Lichtner	Continuum formulation of multicomponent-multiphase reactive transport (1996) [348]	General theory of reactive transport processes in porous media
R. De Boer	Theory of porous media (2000) [114]	Describes historical progression and fundamental equations in a geotechnical context
G.F. Pinder and W.G. Gray	Essentials of multiphase flow and transport in porous media (2008) [422]	Fundamental concepts that underlie the physics of multiphase flow and transport in porous media

Table 1.2 Standard textbooks (classic, research and engineering work)

Author(s)	Title (year) [reference]	Comments
M. Muskat	The flow of homogeneous fluids through porous media (1937) [381]	Summarizes important analytical solutions
A.E. Scheidegger	The physics of flow through porous media (1957) [459]	Pioneering work in groundwater modeling
P. Ya. Polubarinova-Kochina	Theory of groundwater movement 1962) [426]	The classic analytical approach in groundwater modeling
A. Verruijt	Theory of groundwater flow (1970) [546]	Fundamentals of groundwater flow
J. Bear	Hydraulics of groundwater (1979) [34]	The groundwater engineering book
R.A. Freeze and J.A. Cherry	Groundwater (1979) [171]	A comprehensive presentation of groundwater hydrology
G. De Marsily	Quantitative hydrogeology – groundwater hydrology for engineers (1986) [120]	The modeling standard textbook in a hydrogeologic context
O.D.L. Strack	Groundwater mechanics (1989) [492]	Advanced analytical solutions for groundwater problems
G.I. Barenblatt et al.	Theory of fluid flows through natural rocks (1990) [26]	A systematical treatment of the mathematical theory of fluid flows in natural reservoirs
D.A. Nield and A. Bejan	Convection in porous media (2006) [389]	Focusing on fluid-density driven convection processes in porous media
A.T. Corey	Mechanics of immiscible fluids in porous media (1994) [100]	Basic principles of mechanics of two-phase fluid systems
O. Coussy	Mechanics of porous continua (1995) [106]	A rigorous description of solid mechanics for porous media
M. Kaviany	Principles of heat transfer in porous media (1995) [305]	Fundamentals of heat transfer in porous media
J.S. Selker et al.	Vadose zone processes (1999) [473]	An introduction to flow and transport in the vadose (unsaturated) zone
J.W. Delleur (ed.)	The handbook of groundwater engineering (1999) [119]	Covering all important aspects of groundwater modeling in an engineering context
K. Vafai (ed.)	Handbook of porous media (2005) [534]	Modeling flow, heat and mass transport in porous media outside the hydrogeologic and subsurface water context. It also includes non-Darcy flow, convection phenomena, turbulence, combustion and molding processing

Table 1.3 Selected books on finite elements and related numerical techniques (solid and fluid mechanics in general, no specific emphasis on porous-media simulation)

Author(s)	Title (year) [reference]	Comments
G. Strang and G. Fix	An analysis of the finite element method (1973) [493]	Classic book on the mathematics of the FEM
V. Girault and P.A. Raviart	Finite element methods for Navier-Stokes equations. Theory and algorithms (1986) [192]	Mathematical textbook on FEM, a standard reference for the theory of FEM with emphasis on Navier-Stokes equations
C.A.J. Fletcher	Computational techniques for fluid dynamics (1988) [165]	Finite difference, finite elements, finite volume and spectral methods with emphasis on CFD problems
O. Pironneau	Finite element methods for fluids (1989) [423]	Addressing FEM for a wide range of fluid flow problems
O.C. Zienkiewicz and R.L. Taylor	The finite element method: vol. 1 The basis, vol. 2 Solid and structural mechanics, vol. 3 Fluid dynamics (2000) (2002) [590–592]	The 'standards' on finite elements: Giving a broad overview of FEM
P.G. Ciarlet and J.L. Lions (ed.)	Handbook of numerical analysis – Finite element methods (1991) [84]	Advanced mathematical theory on finite elements including error estimates, mixed and hybrid methods
A.J. Baker	Finite element method (1998) [20]	Developing FE algorithms for CFD problems
F. Brezzi and M. Fortin	Mixed and hybrid finite element methods (1991) [56]	Standard reference for mixed and hybrid methods and their stability
B.A. Finlayson	Numerical methods for problems with moving fronts (1992) [164]	Addressing explicit techniques for solving convection-dominated transport equations
S.C. Brenner and L.R. Scott	The mathematical theory of finite element methods (1994) [55]	Modern mathematical theory of FEM
J.N. Reddy and D.K. Gartling	The finite element method in heat transfer and fluid dynamics (2001) [437]	Applied FEM for many heat and fluid flow problems
R. Löhner	Applied CFD techniques (2001) [353]	Numerical methods in CFD covering a number of interesting topics
J.H. Ferziger and M. Peric	Computational methods for fluid dynamics (1996) [162]	An updated textbook on FVM in CFD
P.M. Gresho and R.L. Sani	Incompressible flow and the finite element method (1998) [209]	The 'state of the art' of finite element modeling of Navier-Stokes and advection problems in fluid mechanics
T.J. Chung	Computational fluid dynamics (2002) [83]	Comprehensive book of CFD describing FDM, FVM and FEM techniques in a fluid dynamics context
I.M. Smith and D.V. Griffiths	Programming the finite element method (2004) [484]	Describing a wide variety of problem solving capabilities for FEM
W.J. Minkowycz, E.M. Sparrow and J.Y. Murthy (ed.)	Handbook of numerical heat transfer (2006) [374]	Coverage of formulations, numerical schemes and solution techniques for solving problems of heat and mass transfer and related fluid flows

Table 1.4 Selected books on numerical modeling of subsurface and porous-media problems

Author(s)	Title (year) [reference]	Comments
P.S. Huyakorn and G.F. Pinder	Computational methods in subsurface flow (1983) [280]	A first comprehensive book on modeling of subsurface flow, mass and energy transport with emphasis on FEM
W. Kinzelbach	Groundwater modeling (1986) [310]	An introduction to the major techniques used in modeling groundwater flow and pollutant transport in groundwater. It covers concepts, computational methods and sample programs
M.P. Anderson and W.W. Woessner	Applied groundwater modeling (1992) [9]	Addressed to practiced modelers in groundwater flow and subsurface contaminant transport with emphasis on classic FDM
C. Zheng and G.D. Bennett	Applied contaminant transport modeling (1995) [587]	Describes basic principles of solute transport simulation for porous-media problems. Different numerical approaches are discussed, including particle tracking techniques, FDM, FEM and Lagrangian methods
R. Helmig	Multiphase flow and transport processes in the subsurface (1997) [238]	Useful introductory text on immiscible multiphase process modeling. It covers fundamentals and numerical approaches with FDM, FEM and FVM
E. Holzbecher	Modeling density-driven flow in porous media (1998) [255]	Focusing on modeling variable-density flow. It contains fundamental work and describes streamfunction-related FDM restricted, however, to 2D problems
G.-T. Yeh	Computational subsurface hydrology. Part 1. Fluid flows, Part 2. Reaction, transport, and fate (1999) (2000) [579, 580]	Useful and comprehensive text on numerical subsurface modeling with emphasis on FEM in 2D and 3D applications. Part 2 focuses on reactive geochemical and biochemical transport in porous media
O. Kolditz	Computational methods in environmental fluid mechanics (2002) [317]	Giving an overview on development and application of numerical methods in porous and fractured media. Topics cover fundamental principles, software engineering, flow in fractured media, heat transport in hot dry rock systems, density-dependent flows and deformable porous media
J. Bear and A.H.-D. Cheng	Modeling groundwater flow and contaminant transport (2010) [38]	An updated text on groundwater modeling covering many aspects of model development, single and multiple species transport, FVM, FEM, seawater intrusion, uncertainty, optimization and inverse problems

Monod and arbitrary type are presented. Appropriate initial, boundary and constraint conditions complete the model formulations in Chap. 6. Chapter 7 deals with anisotropy for two- and three-dimensional problems.

Part II of the book describes the basic concept of finite element formulations for solving flow, mass and heat transport in porous and fractured media. It begins with Chap. 8 in which fundamental aspects of the finite element method are thoroughly discussed for prototypical advection-dispersion equations. Weak forms, approaches for spatial and temporal discretization, approximation errors, stability properties, upwinding schemes, treatment of nonlinearities and derived quantities, budget evaluation and local conservativity are reviewed. Based on the general model equations for multispecies, chemically reactive, variable-density, non-Darcy and non-isothermal flow and transport processes in porous and fractured media, their finite element solutions for subclasses of problems are discussed in Chaps. 9–14. They are covered through a number of applications in form of test cases and benchmark examples to examine the presented finite element approaches in comparison to analytical or other numerical solutions. The broad coverage of finite element modeling is provided in Chap. 9 devoting to flow in saturated porous media (groundwater flow), Chap. 10 focusing on flow in variably saturated porous media, Chap. 11 for variable-density flow, mass and heat transport in porous media, Chap. 12 dealing with mass transport in porous media with and without chemical reactions, Chap. 13 referring to heat transport in porous media and Chap. 14 dealing with discrete feature modeling of flow, mass and heat transport processes. The final Chap. 15 discusses specific topics important for the present modeling strategies such as mesh generation, including adaptive mesh refinement/derefinement methods, particle tracking techniques, streamline integration and related finite-element interpolation schemes.

Finally, a number of useful Appendices (A–M) cover tables of used symbols with their physical dimensions, tables of essential parameter relations used for solving the governing flow, mass and heat transport equations, explain mathematical details and summarize important physical relationships.

Part I
Fundamentals

Part I describes the conceptual and physical fundamentals for the modeling of flow, mass and heat transport in porous and fractured media. The derivation of the governing balance equations, phenomenological laws and constitutive relations follows a modern conception based on general principles of continuum mechanics and rational thermodynamics. While such a theoretical framework like this is not really new, it is still rather unusual and underestimated in geosciences and related disciplines, both in education and in modeling practice. However, we believe (and wish to suggest it as an aspirable methodology) that this theoretical development is most physically transparent and clearly different from the traditional level of empiricism we refuse. It provides a widely conflict-free and rigorous derivation of all relevant relationships we need to cover the complete fields of modeling for today's and future applications ranging from most complex multiphase-multispecies flow systems with their convoluted physics and constitutive relations (Darcy-Brinkman-Forchheimer (DBF) flow, non-Fickian dispersion, total energy conservation, cross effects, chemical reactions, solid deformation, extended material laws, density coupling, …) up to the standard porous-medium flow equations governed by the well-known Darcy law. In this process the inherent assumptions for each level of model complexity/simplicity are clearly revealed and become assessable.

In doing so, our preferred methodology for deriving the fundamental model equations is *deductive* per se, where the theory from the most general to the simplest level is developed in a systematic and physically consistent way. Thermodynamic principles are required to constitute the relevant material relationships. We shall recognize, e.g., why phenomenological laws must have negative signs, learn that the Darcy law is only a special case of the momentum conservation of fluid in a porous medium and find that the temperature as the primary variable usually applied to the energy conservation equation for a porous medium is associated with a number of important assumptions. The theoretical developments end up with three levels of model reduction which finally form the actually tractable sets of governing equations summarized in Tables 3.5, 3.7 and 3.9, respectively: (1) multiphase variable-density DBF flow, reactive multispecies mass and heat transport, (2) single-liquid phase variable-density Darcy-type flow, reactive multispecies mass and heat transport in variably saturated porous media, and (3) variable-density Darcy-type flow, reactive multispecies mass and heat transport in groundwater (saturated porous media). Furthermore, aquifer-related model equations are deduced and listed in Tables 3.10 and 3.11 for unconfined and confined conditions, respectively.

At a first glance this deductive development seems fairly extensive. However, the theoretical generalization is required to derive model equations and constitutive relations for the desired spectrum of applications providing different degrees of physical detail and scale. In a practical approach, the reader could study the theoretical developments even in reverse order, where, starting from a standard set of equations, the next higher level of model generalization is examined and the employed steps of assumption can be pursued.

Essential flow and transport equations are also derived for discrete features, which are separated from the porous-medium approach. Discrete features are very useful to model flow and transport processes in fractures, conduits, channels, faults, boreholes and many other macroscopic geometric representations. Typically, diffusion-type flow conditions are assumed in those discrete features. The developments are summarized in Tables 4.5–4.7 for flow, mass and heat transport, respectively, in discrete features.

A comprehensive discussion is presented for chemical reactions, both for reversible and irreversible reaction processes. The developments cover adsorption relations of Henry, Langmuir and Freundlich type as well as reaction kinetics of degradation, Arrhenius and Monod type, including serial-parallel decay, Michaelis-Menten mechanism and freely editable kinetic expressions.

Initial, boundary and constraint conditions are thoroughly discussed for flow, mass and heat transport. Required special formulations of boundary conditions refer to free-surface, seepage-face, surface ponding, integral, gradient-type, multi-layer well and outflow conditions. It is shown that a Neumann-type boundary condition of the divergence form of a transport equation is equivalent to a Cauchy-type boundary condition of its convective form, which easily allows to impose load conditions for mass and heat.

Anisotropy is described in full three dimensions and two dimensions. Important special cases are developed for the shape-derived 3D anisotropy and axis-parallel anisotropy.

Chapter 2
Preliminaries

2.1 Mathematical Foundation

2.1.1 Notation Rules and Algebra with Vectors and Tensors

In the mathematical formulation of quantities and fields there are two types of notation which we shall use through the book: *symbolic (or Gibbs') notation* as well as *index notation*. For the sake of convenience and to provide suitably abstract formulations we mostly prefer the symbolic notation. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be vectors in a (real) space of dimension D, we write with (all used symbols are summarized in the Appendix A)

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_D \end{pmatrix} = a_i \,, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{pmatrix} = b_i \quad (1 \leq i \leq D) \tag{2.1}$$

for the *scalar (or dot) product*

$$\boldsymbol{a} \cdot \boldsymbol{b} = a_i b_i \quad (1 \leq i \leq D) \tag{2.2}$$

using *Einstein's summation convention* according to $a_i b_i = \sum_{i=1}^{D} a_i b_i$ in which repeated indices are summed,
for the *vector (or cross) product*

$$\boldsymbol{a} \times \boldsymbol{b} = \varepsilon_{ijk} a_i b_j \boldsymbol{e}_k = \det \begin{pmatrix} a_1 & a_2 & \dots & a_D \\ b_1 & b_2 & \dots & b_D \\ \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_D \end{pmatrix} \quad (1 \leq i, j, k \leq D) \tag{2.3}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_2,

where ε_{ijk} is the *permutation symbol* (also known as the Levi-Civita tensor $\boldsymbol{\varepsilon}$) defined as

$$\varepsilon_{ijk} = \begin{cases} 1 & \text{if } (ijk) \text{ is an even (cyclic) permutation, e.g., } \varepsilon_{123} = \varepsilon_{231} = \varepsilon_{312} = 1 \\ -1 & \text{if } (ijk) \text{ is an odd (noncyclic) permutation, e.g., } \varepsilon_{213} = \varepsilon_{321} = \varepsilon_{132} = -1 \\ 0 & \text{if two or more subscripts of } (ijk) \text{ are the same, e.g., } \varepsilon_{111} = \varepsilon_{112} = \varepsilon_{313} = 0 \end{cases} \tag{2.4}$$

and $\boldsymbol{e}_i$ $(1 \le i \le D)$ are *base vectors* given as

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad \boldsymbol{e}_D = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \tag{2.5}$$

where

$$\boldsymbol{e}_i \cdot \boldsymbol{e}_j = \boldsymbol{\delta} = \delta_{ij} \tag{2.6}$$

with the *Kronecker symbol (unit or identity matrix)*

$$\boldsymbol{\delta} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \delta_{ij} = \begin{cases} 1 & \text{when} \quad i = j \\ 0 & \text{when} \quad i \neq j \end{cases} \tag{2.7}$$

and for the *dyadic (or tensor) product*

$$\boldsymbol{a} \otimes \boldsymbol{b} = a_i b_j = \begin{pmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_D \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_D \\ \vdots & \vdots & \ddots & \vdots \\ a_D b_1 & a_D b_2 & \cdots & a_D b_D \end{pmatrix} \quad (1 \le i, j \le D) \tag{2.8}$$

which results in a second-order tensor $\boldsymbol{A} = \boldsymbol{a} \otimes \boldsymbol{b}$. The multiplication symbol $\otimes$ in the dyadic product is often omitted and the tensor product of (2.8) is then simply denoted by $\boldsymbol{A} = \boldsymbol{ab}$. We note that the components $a_1, a_2, \dots a_D$ of $\boldsymbol{a}$ in (2.1) are themselves *scalars* and the vector $\boldsymbol{a}$ can also be formed via summation

$$\boldsymbol{a} = a_i \boldsymbol{e}_i \quad (1 \le i \le D) \tag{2.9}$$

We further note that (2.9) is a symbolic vector expression. In such a context a_i are scalars and not seen as a vector symbol used in the index notation. An equivalent expression for (2.9) reads $a_i = \delta_{ij} a_j$ in the index notation. In generalization, the following convention for vector multiplication holds in symbolic notation:

$$\begin{array}{ll} a_i \boldsymbol{e}_i & \text{results in a vector, where } a_i \text{ represent vector components (scalars),} \\ (a_i) \cdot \boldsymbol{e}_i & \text{results in a scalar, where } (a_i) \text{ represents a vector,} \\ (a_{ij}) \cdot \boldsymbol{e}_j & \text{results in a vector, where } (a_{ij}) \text{ represents a second-order tensor.} \end{array} \tag{2.10}$$

The *norm (or magnitude) a* of vector $\boldsymbol{a}$ is given by

$$a = \|\boldsymbol{a}\| = \sqrt{\boldsymbol{a} \cdot \boldsymbol{a}} = \sqrt{a_i a_i} \tag{2.11}$$

If $\|\boldsymbol{a}\| = 1$ it is called a *unit vector* such as $\boldsymbol{e}_i$.
Furthermore, we can find the *normalized vector* for $\boldsymbol{a}$ according to

$$\hat{\boldsymbol{a}} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} = \frac{a_i}{\sqrt{a_j a_j}} \boldsymbol{e}_i \tag{2.12}$$

It becomes clear that $\hat{\boldsymbol{a}}$ is itself a unit vector because $\|\hat{\boldsymbol{a}}\| = 1$. The *transpose* of vector $\boldsymbol{a}$ changes a column vector to a row vector and a row vector to a column vector, respectively,

$$\boldsymbol{a}^T = \begin{pmatrix} a_1 \; a_2 \dots a_D \end{pmatrix} \qquad \begin{pmatrix} a_1 \; a_2 \dots a_D \end{pmatrix}^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_D \end{pmatrix} \tag{2.13}$$

Let $\boldsymbol{A} = \sum_i \sum_j A_{ij} \boldsymbol{e}_i \otimes \boldsymbol{e}_j$ and $\boldsymbol{B} = \sum_i \sum_j B_{ij} \boldsymbol{e}_i \otimes \boldsymbol{e}_j$ be two second-order tensors of dimension D $(1 \leq i, j \leq D)$

$$\boldsymbol{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1D} \\ A_{21} & A_{22} & \cdots & A_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ A_{D1} & A_{D2} & \cdots & A_{DD} \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1D} \\ B_{21} & B_{22} & \cdots & B_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ B_{D1} & B_{D2} & \cdots & B_{DD} \end{pmatrix} \tag{2.14}$$

then, their scalar product is written as a *double dot product* (or colon product) in the form

$$\boldsymbol{A} : \boldsymbol{B} = \sum_i^D \sum_j^D A_{ij} B_{ji} \tag{2.15}$$

which results in a scalar. The norm of such a second-order tensor $\boldsymbol{A}$ is defined as

$$\|\boldsymbol{A}\| = \sqrt{\boldsymbol{A} : \boldsymbol{A}^T} = \sqrt{\sum_i \sum_j (A_{ij})^2} \tag{2.16}$$

where the transpose $\boldsymbol{A}^T$ of the tensor $\boldsymbol{A}$ is given by $(A_{ij})^T = (A_{ji})$:

$$\boldsymbol{A}^T = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{D1} \\ A_{12} & A_{22} & \cdots & A_{D2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1D} & A_{2D} & \cdots & A_{DD} \end{pmatrix} \tag{2.17}$$

A second-order tensor $\boldsymbol{A} = A_{ij}$ is a *symmetric tensor* for which the following is valid

$$\boldsymbol{A} = \boldsymbol{A}^T \qquad A_{ij} = A_{ji} \tag{2.18}$$

Any tensor $\boldsymbol{A}$ can be written as a sum of symmetric and antisymmetric parts

$$\boldsymbol{A} = \tfrac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^T) + \tfrac{1}{2}(\boldsymbol{A} - \boldsymbol{A}^T) = \tfrac{1}{2}(\boldsymbol{B}_s + \boldsymbol{B}_a) \tag{2.19}$$

The scalar product of a tensor $\boldsymbol{A}$ with a vector $\boldsymbol{a}$ is:

$$\boldsymbol{A} \cdot \boldsymbol{a} = \sum_i^D \boldsymbol{e}_i \sum_j^D A_{ij} a_j \tag{2.20}$$

In contrast, the scalar product of a vector $\boldsymbol{a}$ with a tensor $\boldsymbol{A}$ is:

$$\boldsymbol{a} \cdot \boldsymbol{A} = \sum_i^D \boldsymbol{e}_i \sum_j^D a_j A_{ji} \tag{2.21}$$

A tensor $\boldsymbol{A}$ is *diagonal* if the components outside the main diagonal are all zero, i.e., $A_{ij} = 0$ for $i \neq j$. It is written as

$$\boldsymbol{A} = \lceil A_{11}, A_{22}, \ldots, A_{DD} \rfloor = \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{DD} \end{pmatrix} \tag{2.22}$$

Any diagonal tensor is also a symmetric tensor.

2.1.2 Relations for Scalar and Vector Products

If $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ are unit vectors in the directions of $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively, then

$$\hat{\boldsymbol{a}} \cdot \hat{\boldsymbol{b}} = \cos\theta \tag{2.23}$$

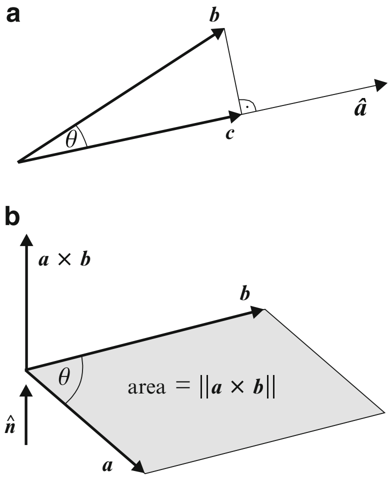


Fig. 2.1 (**a**) Projection of $\boldsymbol{b}$ onto the unit vector $\hat{\boldsymbol{a}}$ and (**b**) vector product $\boldsymbol{a} \times \boldsymbol{b}$

where θ represents the angle between the two directions. Since $\boldsymbol{a} = \|\boldsymbol{a}\|\,\hat{\boldsymbol{a}}$ and $\boldsymbol{b} = \|\boldsymbol{b}\|\,\hat{\boldsymbol{b}}$ we find for the scalar product (2.2) with (2.23)

$$\boldsymbol{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|\,\cos\theta \tag{2.24}$$

This form of the scalar product is useful to formulate the *projection* of a vector onto a given direction. Assuming we have a unit vector $\hat{\boldsymbol{a}}$ and another vector $\boldsymbol{b}$, we project $\boldsymbol{b}$ perpendicularly onto $\hat{\boldsymbol{a}}$, as shown in Fig. 2.1, and call the resulting projected vector $\boldsymbol{c}$. We find

$$\begin{aligned} \|\boldsymbol{c}\| &= \|\boldsymbol{b}\|\,\cos\theta \\ &= \|\boldsymbol{b}\|\Big(\frac{\hat{\boldsymbol{a}} \cdot \boldsymbol{b}}{\|\hat{\boldsymbol{a}}\|\,\|\boldsymbol{b}\|}\Big) \\ &= \hat{\boldsymbol{a}} \cdot \boldsymbol{b} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\|} \end{aligned} \tag{2.25}$$

The interpretation of (2.25) is that the scalar product of the unit vector in the direction of vector $\boldsymbol{a}$ and the vector $\boldsymbol{b}$ yields the length of the projection of $\boldsymbol{b}$ onto $\hat{\boldsymbol{a}}$ (see Fig. 2.1). If the angle θ between the two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is a right angle, $\theta = \pi/2$, then $\cos\theta = 0$. Such vectors are said to be orthogonal and the condition for *orthogonality* is

$$\boldsymbol{a} \cdot \boldsymbol{b} = 0 \tag{2.26}$$

The *square* of a vector $\boldsymbol{a}$ results from the scalar product (2.24) with (2.11)

$$\boldsymbol{a} \cdot \boldsymbol{a} = \|\boldsymbol{a}\| \|\boldsymbol{a}\| \cos 0 = a^2 \tag{2.27}$$

The cross product $\boldsymbol{a} \times \boldsymbol{b}$ of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ forms a vector of magnitude $\|\boldsymbol{a}\| \; \|\boldsymbol{b}\| \; \sin\theta$ normal to the plane defined by $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\boldsymbol{a} \times \boldsymbol{b} = \|\boldsymbol{a}\| \; \|\boldsymbol{b}\| \; \sin\theta \; \hat{\boldsymbol{n}} \tag{2.28}$$

where the unit vector $\hat{\boldsymbol{n}}$ is normal to the plane $\boldsymbol{a}$ and $\boldsymbol{b}$ and $\|\boldsymbol{a} \times \boldsymbol{b}\|$ is the area of the parallelogram that the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ span.

2.1.3 Coordinate System and Spatial Vector

To position physical objects in space and to define their spatial motion, the D-dimensional Euclidean space $\Re^D$ $(D = 1, 2, 3)$ is used as reference system. We employ an orthogonal Cartesian coordinate system as shown in Fig. 2.2 in which a position P is defined by the Cartesian coordinate vector $\boldsymbol{x}$, viz.,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \text{in } \Re^3 \text{ (3D)} \tag{2.29}$$

In $\Re^2$ (2D) and $\Re^1$ (1D) it is $\boldsymbol{x}^T = (x_1 \;\; x_2) = (x \;\; y)$ and $\boldsymbol{x}^T = (x_1) = (x)$, respectively. If we want to identify positions of a list of points $\boldsymbol{x}_l \in \Re^D$ $(l = 1, 2, \ldots, N_\mathrm{P})$ labeled by a (nodal) index l, their coordinates are written as

$$\boldsymbol{x}_l = \begin{pmatrix} x_{1l} \\ x_{2l} \\ x_{3l} \end{pmatrix} = \begin{pmatrix} x_l \\ y_l \\ z_l \end{pmatrix} \quad \text{in } \Re^3 \text{ (3D)}, \quad (l = 1, 2, \ldots, N_\mathrm{P}) \tag{2.30}$$

where N_P is the number of the listed points. It follows that the components of the coordinate vector are themselves vectors consisting of an ordered list of (discrete) numbers, i.e., $\boldsymbol{x}_1^T = (x_{11} \;\; x_{12} \;\; \ldots \;\; x_{1N_\mathrm{P}})$, and so forth. With the coordinate vector $\boldsymbol{x} = x_i \boldsymbol{e}_i$ we find that the corresponding base vectors $\boldsymbol{e}_i$ are formed from the tangent vectors $\bar{\boldsymbol{e}}_i$ to the coordinate lines of the $\boldsymbol{x}-$coordinates, viz.,

$$\begin{aligned} \boldsymbol{e}_i &= \widehat{\bar{\boldsymbol{e}}}_i = \frac{\bar{\boldsymbol{e}}_i}{\|\bar{\boldsymbol{e}}_i\|} \\ \bar{\boldsymbol{e}}_i &= \frac{\partial \boldsymbol{x}}{\partial x_i} \end{aligned} \tag{2.31}$$

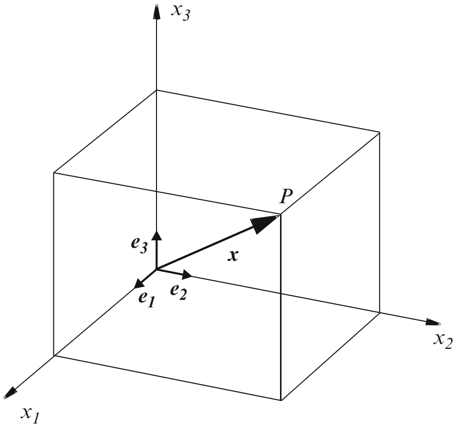


Fig. 2.2 Cartesian coordinate system in $\Re^3$. The position P is represented by the vector $\boldsymbol{x} = x_i \boldsymbol{e}_i$

We note that the Cartesian coordinate system is orthogonal with $\boldsymbol{e}_i \cdot \boldsymbol{e}_j = 0 \, (i \neq j)$, cf. (2.26).

2.1.4 Eulerian and Lagrangian Coordinates

Following the concepts of continuum mechanics we use the term *point* to indicate a location in space $\Re^D$ and the term *particle* to denote a *point in a continuum*. While points are fixed in space and independent of time t, positions of particles may vary with time t. This provides a distinction between two kinds of coordinates:

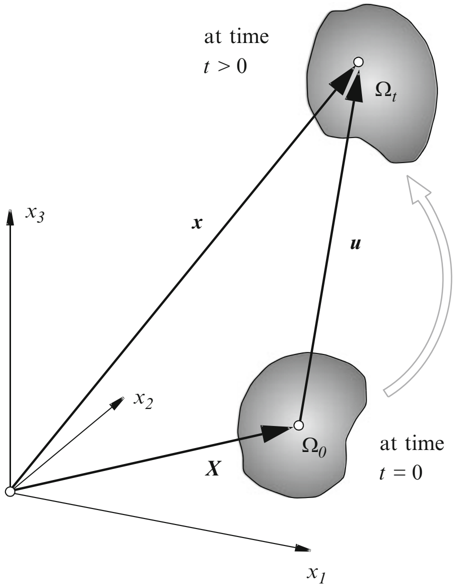
(a) *Spatial (Eulerian) coordinates* $\boldsymbol{x}$, that define points in space with respect to a fixed frame of reference.
(b) *Material (Lagrangian) coordinates* $\boldsymbol{X}$, that are assigned to particles of a continuum. Usually, $\boldsymbol{X}$ is selected as the initial position vector of a considered particle, i.e., $\boldsymbol{X} \equiv \boldsymbol{x}|_{t=0}$.

As a particle moves, its coordinates $\boldsymbol{x}$ vary in time t, whereas its material coordinates $\boldsymbol{X}$ remain unchanged. Such a motion is described by

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{X}, t) \tag{2.32}$$

which is known as the *Lagrangian formulation of motion*. Figure 2.3 shows a spatial domain Ω_0 occupied at $t = 0$ by a continuum with material coordinates $\boldsymbol{X}$. At a later time $t > 0$, the domain occupied by the same continuum is Ω_t. The domain Ω_t represents the *deformed configuration* of the continuum initially in Ω_0.

Fig. 2.3 Motion of a particle



The *Eulerian formulation of motion* is obtained if (2.32) is inverted to yield the initial position (i.e., material coordinates) of a particle which at time t is at position $\boldsymbol{x}$:

$$\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{x}, t) \tag{2.33}$$

A necessary and sufficient condition for the existence of (2.33) is given, if the Jacobian

$$\boldsymbol{J} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{X}} = \begin{pmatrix} \dfrac{\partial x_1}{\partial X_1} & \dfrac{\partial x_2}{\partial X_1} & \dfrac{\partial x_3}{\partial X_1} \\ \dfrac{\partial x_1}{\partial X_2} & \dfrac{\partial x_2}{\partial X_2} & \dfrac{\partial x_3}{\partial X_2} \\ \dfrac{\partial x_1}{\partial X_3} & \dfrac{\partial x_2}{\partial X_3} & \dfrac{\partial x_3}{\partial X_3} \end{pmatrix} \tag{2.34}$$

differs from zero. The *displacement* vector $\boldsymbol{u}$ (Fig. 2.3) is defined as the difference between the position vector $\boldsymbol{x}$ of a moving particle at a given time t and its initial position vector $\boldsymbol{X}$:

$$\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{X} \tag{2.35}$$

The need for a mathematical description based on a fixed domain and spatial reference renders the Eulerian formulation an ideal candidate to describe flow fields.

Accordingly, the Eulerian concept is preferred in our analysis. It requires, however, that the flow and transport quantities be *continuous* throughout the domain Ω. Let us consider a property $f = f(\boldsymbol{x}, t)$ in the Eulerian description, which is linked to the considered particle in Ω. Using (2.32) we obtain the relationship between the two respective coordinate systems

$$f(\boldsymbol{x}, t) = f[\boldsymbol{x}(\boldsymbol{X}, t), t] \tag{2.36}$$

and the rate of change of f in a Lagrangian description

$$\frac{Df}{Dt} = \left.\frac{\partial f}{\partial t}\right|_{\boldsymbol{X}=\text{const}} \tag{2.37}$$

defining the derivative with respect to time t keeping $\boldsymbol{X}$ constant. The derivative D/Dt represents the rate of change as observed when moving with the particle and is called *material derivative*. In the Eulerian description we can derive

$$\begin{aligned}\frac{Df[\boldsymbol{x}(\boldsymbol{X}, t), t]}{Dt} &= \left.\frac{\partial f}{\partial t}\right|_{\boldsymbol{x}=\text{const}} + \left.\frac{\partial f}{\partial \boldsymbol{x}}\right|_{t=\text{const}} \cdot \left.\frac{\partial \boldsymbol{x}(\boldsymbol{X}, t)}{\partial t}\right|_{\boldsymbol{X}=\text{const}} \\ &= \frac{\partial f}{\partial t} + (\boldsymbol{v} \cdot \nabla) f\end{aligned} \tag{2.38}$$

where

$$\boldsymbol{v} = \dot{\boldsymbol{x}} = \left.\frac{\partial \boldsymbol{x}}{\partial t}\right|_{\boldsymbol{X}=\text{const}} \tag{2.39}$$

is the *velocity* of the particle and

$$\nabla = \frac{\partial}{\partial \boldsymbol{x}} = \boldsymbol{e}_i \frac{\partial}{\partial x_i} \tag{2.40}$$

is the *gradient (or Nabla) operator* which represents a $D-$dimensional vector.

2.1.5 Coordinate Transformations

Physical quantities in form of scalars, vectors and tensors have to be coordinate-invariant properties. In transforming between different coordinate systems the quantities have to remain unchanged. It is important to determine the relations between sets of components relative to different coordinate systems. The introduction of different coordinate systems is often useful to simplify the analysis. In our applications we need to use *orthogonal coordinate systems*, i.e., systems where the coordinate lines are orthogonal, for example cylindrical coordinates or local finite-element coordinates.

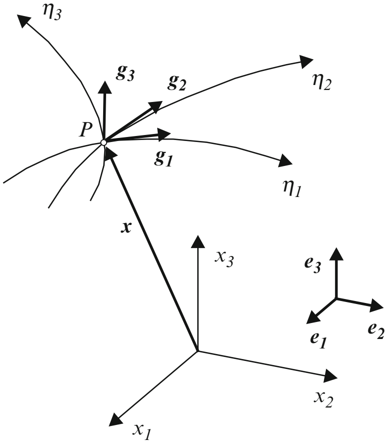


Fig. 2.4 Transformed coordinate system $\boldsymbol{\eta}$

2.1.5.1 Mapping

Introducing the general coordinates

$$\begin{aligned}\boldsymbol{\eta} &= \begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix}\\ \boldsymbol{x} &= \eta_i \boldsymbol{g}_i\end{aligned} \tag{2.41}$$

where $\boldsymbol{g}_i$ are the base vectors of the $\boldsymbol{\eta}-$system (Fig. 2.4), a one-to-one mapping between the $\boldsymbol{\eta}-$space and the Euclidean $\boldsymbol{x}-$space must exist:

$$\boldsymbol{x}(\boldsymbol{\eta}) \Leftrightarrow \boldsymbol{\eta}(\boldsymbol{x}) \tag{2.42}$$

The corresponding tangent vectors $\bar{\boldsymbol{g}}_i$ (2.41) of the $\boldsymbol{\eta}-$coordinates provide, cf. (2.31)

$$\bar{\boldsymbol{g}}_i = \frac{\partial \boldsymbol{x}}{\partial \eta_i} = \frac{\partial \boldsymbol{x}}{\partial x_k} \cdot \left(\frac{\partial x_k}{\partial \eta_i}\right) = \left(\frac{\partial x_k}{\partial \eta_i}\right) \cdot \bar{\boldsymbol{e}}_k = (J_{ik}) \cdot \bar{\boldsymbol{e}}_k \tag{2.43}$$

where the Jacobian matrix $\boldsymbol{J}$

$$\boldsymbol{J} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\eta}} = J_{ij} = \frac{\partial x_j}{\partial \eta_i} = \begin{pmatrix} \frac{\partial x_1}{\partial \eta_1} & \frac{\partial x_2}{\partial \eta_1} & \frac{\partial x_3}{\partial \eta_1}\\ \frac{\partial x_1}{\partial \eta_2} & \frac{\partial x_2}{\partial \eta_2} & \frac{\partial x_3}{\partial \eta_2}\\ \frac{\partial x_1}{\partial \eta_3} & \frac{\partial x_2}{\partial \eta_3} & \frac{\partial x_3}{\partial \eta_3} \end{pmatrix} = \begin{pmatrix} J_{11} & J_{12} & J_{13}\\ J_{21} & J_{22} & J_{23}\\ J_{31} & J_{32} & J_{33} \end{pmatrix} \tag{2.44}$$

must be non-zero to perform the reverse transformation

$$\bar{\boldsymbol{e}}_k = \left(\frac{\partial \eta_i}{\partial x_k}\right) \cdot \bar{\boldsymbol{g}}_i = (J_{ki})^{-1} \cdot \bar{\boldsymbol{g}}_i \tag{2.45}$$

We note that the base vectors are derived from the tangent vectors, viz.,

$$\boldsymbol{g}_i = \frac{\bar{\boldsymbol{g}}_i}{\|\bar{\boldsymbol{g}}_i\|} \qquad \boldsymbol{e}_i = \frac{\bar{\boldsymbol{e}}_i}{\|\bar{\boldsymbol{e}}_i\|} \tag{2.46}$$

With $\boldsymbol{a}^x$ denoting an arbitrary vector in the $\boldsymbol{x}-$system, the corresponding vector $\boldsymbol{a}^\eta$ expressed in the transformed $\boldsymbol{\eta}-$system can be obtained via projection, cf. (2.25)

$$\boldsymbol{a}^\eta = \begin{pmatrix} \boldsymbol{a}^x \cdot \boldsymbol{g}_1 \\ \boldsymbol{a}^x \cdot \boldsymbol{g}_2 \\ \boldsymbol{a}^x \cdot \boldsymbol{g}_3 \end{pmatrix} \tag{2.47}$$

According to (2.45) a Cartesian vector $\boldsymbol{a}^x$ can be expressed by the vector components a_i^η written in the $\boldsymbol{\eta}-$coordinate system:

$$\boldsymbol{a}^x = \begin{pmatrix} a_1^x \\ a_2^x \\ a_3^x \end{pmatrix} = \begin{pmatrix} (J_{1i})^{-1}\, a_i^\eta \\ (J_{2i})^{-1}\, a_i^\eta \\ (J_{3i})^{-1}\, a_i^\eta \end{pmatrix} \qquad a_i^\eta = \begin{pmatrix} a_1^\eta \\ a_2^\eta \\ a_3^\eta \end{pmatrix} = \boldsymbol{a}^\eta \tag{2.48}$$

2.1.5.2 Cylindrical Coordinate System

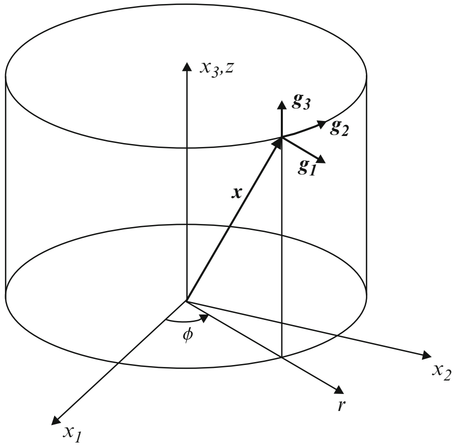
We apply *cylindrical coordinates* $\boldsymbol{\eta}^T = (r \;\; \phi \;\; z)$, where the mapping $\boldsymbol{x}(\boldsymbol{\eta})$ is given by

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} r\cos\phi \\ r\sin\phi \\ z \end{pmatrix} \tag{2.49}$$

in which (r, ϕ, z) correspond to the radial, azimuthal and axial coordinates, respectively (see Fig. 2.5). For cylindrical coordinates the Jacobian $\boldsymbol{J}$ yields

$$\boldsymbol{J} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -r\sin\phi & r\cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \boldsymbol{J}^{-1} = \frac{1}{|\boldsymbol{J}|}\begin{pmatrix} r\cos\phi & -\sin\phi & 0 \\ r\sin\phi & \cos\phi & 0 \\ 0 & 0 & r \end{pmatrix} \qquad |\boldsymbol{J}| = r \tag{2.50}$$

Fig. 2.5 Cylindrical coordinates (r, ϕ, z)



where the determinant of matrix $\boldsymbol{J}$ is denoted by $|\boldsymbol{J}| = \det \boldsymbol{J}$.According to (2.43) and (2.50) we obtain the base vectors $\boldsymbol{g}_i$ to express cylindrical coordinates via Cartesian coordinates

$$\begin{aligned}
\boldsymbol{g}_1^T &= \frac{(J_{11}\ J_{12}\ J_{13})}{\|(J_{11}\ J_{12}\ J_{13})\|} = (\cos\phi\ \ \sin\phi\ 0) \\
\boldsymbol{g}_2^T &= \frac{(J_{21}\ J_{22}\ J_{23})}{\|(J_{21}\ J_{22}\ J_{23})\|} = (-\sin\phi\ \ \cos\phi\ 0) \\
\boldsymbol{g}_3^T &= \frac{(J_{31}\ J_{32}\ J_{33})}{\|(J_{31}\ J_{32}\ J_{33})\|} = (0\ 0\ 1)
\end{aligned} \tag{2.51}$$

It can easily be shown that the spatial derivatives with respect to the cylindrical coordinates such as $\frac{\partial \boldsymbol{g}_i}{\partial r}$, $\frac{\partial \boldsymbol{g}_i}{\partial \phi}$ and $\frac{\partial \boldsymbol{g}_i}{\partial z}$ become zero, except for

$$\begin{aligned}
\frac{\partial \boldsymbol{g}_1^T}{\partial \phi} &= (-\sin\phi\ \ \cos\phi\ 0) = \boldsymbol{g}_2^T \\
\frac{\partial \boldsymbol{g}_2^T}{\partial \phi} &= (-\cos\phi\ \ -\sin\phi\ 0) = -\boldsymbol{g}_1^T
\end{aligned} \tag{2.52}$$

Exemplified for the Nabla operator (2.40) we can find the transformation by using (2.48) and (2.50):

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} \end{pmatrix} = \begin{pmatrix} \cos\phi \frac{\partial}{\partial r} - \frac{1}{r}\sin\phi \frac{\partial}{\partial \phi} \\ \sin\phi \frac{\partial}{\partial r} + \frac{1}{r}\cos\phi \frac{\partial}{\partial \phi} \\ \frac{\partial}{\partial z} \end{pmatrix} \tag{2.53}$$

Taking into account (2.47), the Nabla operator in the $\boldsymbol{\eta}-$system is built from $(\nabla_i)\cdot \boldsymbol{g}_i$ and finally results with (2.53) and (2.51) in

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{1}{r}\frac{\partial}{\partial \phi} \\ \frac{\partial}{\partial z} \end{pmatrix} \tag{2.54}$$

Let $\boldsymbol{a}$ be a vector in cylindrical coordinates $\boldsymbol{a}^T = (a_r \; a_\phi \; a_z)$, then the scalar product $\nabla \cdot \boldsymbol{a}$ results in cylindrical coordinates using (2.54) and (2.52):

$$\begin{aligned} \nabla \cdot \boldsymbol{a} &= (\boldsymbol{g}_1 \tfrac{\partial}{\partial r} + \boldsymbol{g}_2 \tfrac{1}{r}\tfrac{\partial}{\partial \phi} + \boldsymbol{g}_3 \tfrac{\partial}{\partial z}) \cdot (\boldsymbol{g}_1 a_r + \boldsymbol{g}_2 a_\phi + \boldsymbol{g}_3 a_z) \\ &= \frac{1}{r}\frac{\partial (r\, a_r)}{\partial r} + \frac{1}{r}\frac{\partial a_\phi}{\partial \phi} + \frac{\partial a_z}{\partial z} \end{aligned} \tag{2.55}$$

In the same way we find for the vector product $\nabla \times \boldsymbol{a}$ in cylindrical coordinates

$$\nabla \times \boldsymbol{a} = \begin{pmatrix} \frac{1}{r}\frac{\partial a_z}{\partial \phi} - \frac{\partial a_\phi}{\partial z} \\ \frac{\partial a_r}{\partial z} - \frac{\partial a_z}{\partial r} \\ \frac{1}{r}[\frac{\partial}{\partial r}(r\, a_\phi) - \frac{\partial a_r}{\partial \phi}] \end{pmatrix} \tag{2.56}$$

2.1.5.3 Rotated Coordinate System

Another orthogonal coordinate transformation of interest is the *rotation* of $\boldsymbol{x}-$coordinates in the form

$$\boldsymbol{\eta} = \boldsymbol{A} \cdot \boldsymbol{x} \quad \text{with the rotation matrix} \quad \boldsymbol{A} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \tag{2.57}$$

and, accordingly

$$\boldsymbol{x} = \boldsymbol{A}^{-1} \cdot \boldsymbol{\eta} \quad \text{with} \quad \boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} \begin{pmatrix} A_{22}A_{33} - A_{32}A_{23} & A_{32}A_{13} - A_{12}A_{33} & A_{12}A_{23} - A_{22}A_{13} \\ A_{31}A_{23} - A_{21}A_{33} & A_{11}A_{33} - A_{31}A_{13} & A_{21}A_{13} - A_{11}A_{23} \\ A_{21}A_{32} - A_{31}A_{22} & A_{31}A_{12} - A_{11}A_{32} & A_{11}A_{22} - A_{21}A_{12} \end{pmatrix} \tag{2.58}$$

and

$$|\boldsymbol{A}| = A_{11}(A_{22}A_{33} - A_{32}A_{23}) + A_{21}(A_{32}A_{13} - A_{12}A_{33}) + A_{31}(A_{12}A_{23} - A_{22}A_{13}) \tag{2.59}$$

where A_{ij} are *directional cosines* which are given by

$$A_{ij} \equiv \cos(\boldsymbol{g}_i, \boldsymbol{e}_j) \tag{2.60}$$

For a rotation about the x_3-axis we get (Fig. 2.6)

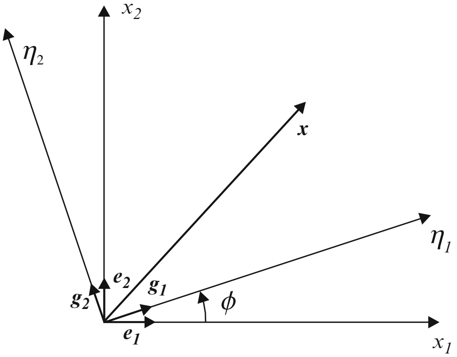


Fig. 2.6 Rotation of coordinates around x_3−axis as a 2D orthogonal coordinate transformation

$$A = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.61}$$

and a full rotation about all three axes in the 3D space the rotation matrix $\boldsymbol{A}$ yields [194]

$$\boldsymbol{A} = \begin{pmatrix} \cos\psi\cos\phi - \cos\theta\sin\phi\sin\psi & \cos\psi\sin\phi + \cos\theta\cos\phi\sin\psi & \sin\psi\sin\theta \\ -\sin\psi\cos\phi - \cos\theta\sin\phi\cos\psi & -\sin\psi\sin\phi + \cos\theta\cos\phi\cos\psi & \cos\psi\sin\theta \\ \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{pmatrix} \tag{2.62}$$

where (ϕ, θ, ψ) are the *Eulerian angles* as defined in Fig. 2.7.

2.1.5.4 Volume, Surface and Line Integral Elements

In Cartesian coordinates the elements of volume $d\Omega$, surface $d\Gamma$ and line dS are

$$\begin{aligned} d\Omega &= dx_1\, dx_2\, dx_3 \\ d\Gamma &= dx_i\, dx_j \qquad (i \neq j, 1 \leq i, j \leq 3) \\ dS &= dx_i \end{aligned} \tag{2.63}$$

Since

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\eta}) \qquad x_i = x_i(\eta_1, \eta_2, \eta_3) \tag{2.64}$$

then by partial differentiation we can derive

$$d\boldsymbol{x} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\eta}} \cdot d\boldsymbol{\eta} = \boldsymbol{J} \cdot d\boldsymbol{\eta} \qquad dx_i = \frac{\partial x_i}{\partial \eta_j} d\eta_j = J_{ij}\eta_j \tag{2.65}$$

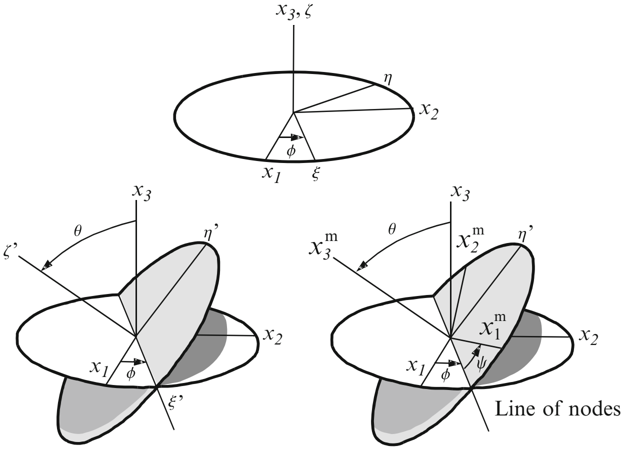


Fig. 2.7 The rotations defining the Eulerian angles in 3D. Note that (ξ, η, ζ) and (ξ', η', ζ') represent intermediate stages of a sequential rotation of the axes (Modified from [194])

Let $d\boldsymbol{x}_i$ be the vectors with components $(\partial x_i / \partial \eta_j) d\eta_j$ for $(j = 1, 2, 3)$, then we obtain for the volume element $d\Omega$

$$\begin{aligned} d\Omega &= (d\boldsymbol{x}_1 \times d\boldsymbol{x}_2) \cdot d\boldsymbol{x}_3 \\ &= |\boldsymbol{J}| d\eta_1\, d\eta_2\, d\eta_3 \end{aligned} \tag{2.66}$$

for the surface element $d\Gamma$ for instance

$$\begin{aligned} d\Gamma &= \|d\boldsymbol{x}_1 \times d\boldsymbol{x}_2\| \\ &= \left\| \left. \begin{pmatrix} J_{12}J_{23} - J_{22}J_{13} \\ J_{21}J_{13} - J_{11}J_{23} \\ J_{11}J_{22} - J_{21}J_{12} \end{pmatrix} \right|_{\eta_3} \right\| d\eta_1\, d\eta_2 \end{aligned} \tag{2.67}$$

and for the line element dS for instance

$$\begin{aligned} dS &= \|d\boldsymbol{x}_1\| \\ &= \left\| \left. \begin{pmatrix} J_{11} \\ J_{12} \\ J_{13} \end{pmatrix} \right|_{\eta_2, \eta_3} \right\| d\eta_1 = \left. \sqrt{J_{11}^2 + J_{12}^2 + J_{13}^2} \right|_{\eta_2, \eta_3} d\eta_1 \end{aligned} \tag{2.68}$$

where $|\boldsymbol{J}|$ is the determinant of the Jacobian (2.44) and $\|.\|$ represents the norm of the vector resulting from transformed coordinates. It is to be noted that $d\Gamma$ and dS can also be expressed by different transformed coordinates, e.g., $d\Gamma = \|(.)|_{\eta_1}\| d\eta_2\, d\eta_3$ and so on. For example, by using cylindrical coordinates (r, ϕ, z) the following integral elements result according to the mapping (2.50):

$$\begin{aligned} d\Omega &= r\,dr\,d\phi\,dz \quad \text{in the } (r,\phi,z) \text{ space} \\ d\Gamma &= \begin{cases} r\,dr\,d\phi & \text{in the } (r,\phi)\text{−space} \\ dr\,dz & \text{in the } (r,z)\text{−space} \\ r\,d\phi\,dz & \text{in the } (\phi,z)\text{−space} \end{cases} \\ dS &= \begin{cases} dr & \text{in the } r\text{−space} \\ r\,d\phi & \text{in the } \phi\text{−space} \\ dz & \text{in the } z\text{−space} \end{cases} \end{aligned} \tag{2.69}$$

2.1.6 Spatial Variables and Their Derivative Operations

For both Cartesian $\Re^D$ $(D = 1, 2, 3)$ and cylindrical coordinate systems

$$\boldsymbol{x}^T = \left\{ \begin{array}{c} (x_1\ x_2\ x_3) \\ (x_1\ x_2) \\ (x_1) \\ (r\ \phi\ z) \end{array} \right\} \text{for} \left\{ \begin{array}{l} \left.\begin{array}{c} \text{3D} \\ \text{2D} \\ \text{1D} \end{array}\right\} \text{Cartesian} \\ \text{cylindrical} \end{array} \right. \tag{2.70}$$

a scalar variable ψ and the velocity $\boldsymbol{v}$ (2.39)

$$\boldsymbol{v} = \begin{cases} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} & \text{for Cartesian} \\ \begin{pmatrix} v_r \\ v_\phi \\ v_z \end{pmatrix} & \text{for cylindrical} \end{cases} \tag{2.71}$$

have dependencies in space and time: $\psi = \psi(\boldsymbol{x}, t)$, $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$. The following derivative operations hold, cf. (2.2), (2.3), (2.40), (2.54), (2.55) and (2.56). The gradient $\nabla\psi$ is

$$\nabla\psi = \begin{cases} \left(\dfrac{\partial\psi}{\partial x_1}\ \dfrac{\partial\psi}{\partial x_2}\ \dfrac{\partial\psi}{\partial x_3}\right)^T & \text{3D Cartesian} \\ \left(\dfrac{\partial\psi}{\partial x_1}\ \dfrac{\partial\psi}{\partial x_2}\right)^T & \text{2D Cartesian} \\ \left(\dfrac{\partial\psi}{\partial x_1}\right) & \text{1D Cartesian} \\ \left(\dfrac{\partial\psi}{\partial r}\ \dfrac{1}{r}\dfrac{\partial\psi}{\partial\phi}\ \dfrac{\partial\psi}{\partial z}\right)^T & \text{cylindrical } (r,\phi,z) \end{cases} \tag{2.72}$$

The second-order derivative (Laplacian) operation $\nabla^2\psi$ reads to

$$\nabla^2\psi = \begin{cases} \dfrac{\partial^2\psi}{\partial x_1^2} + \dfrac{\partial^2\psi}{\partial x_2^2} + \dfrac{\partial^2\psi}{\partial x_3^2} & \text{3D Cartesian} \\ \dfrac{\partial^2\psi}{\partial x_1^2} + \dfrac{\partial^2\psi}{\partial x_2^2} & \text{2D Cartesian} \\ \dfrac{\partial^2\psi}{\partial x_1^2} & \text{1D Cartesian} \\ \dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\,\dfrac{\partial\psi}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial^2\psi}{\partial\phi^2} + \dfrac{\partial^2\psi}{\partial z^2} & \text{cylindrical } (r,\phi,z) \end{cases} \tag{2.73}$$

The scalar product $\nabla\cdot\boldsymbol{v}$ which is called the *divergence* of the vector $\boldsymbol{v}$ is given by

$$\nabla\cdot\boldsymbol{v} = \begin{cases} \dfrac{\partial v_1}{\partial x_1} + \dfrac{\partial v_2}{\partial x_2} + \dfrac{\partial v_3}{\partial x_3} & \text{3D Cartesian} \\ \dfrac{\partial v_1}{\partial x_1} + \dfrac{\partial v_2}{\partial x_2} & \text{2D Cartesian} \\ \dfrac{\partial v_1}{\partial x_1} & \text{1D Cartesian} \\ \dfrac{1}{r}\dfrac{\partial(r\,v_r)}{\partial r} + \dfrac{1}{r}\dfrac{\partial v_\phi}{\partial\phi} + \dfrac{\partial v_z}{\partial z} & \text{cylindrical } (r,\phi,z) \end{cases} \tag{2.74}$$

The vector product $\nabla\times\boldsymbol{v}$ which is called the *curl* of the vector $\boldsymbol{v}$, also known as *vorticity* $\boldsymbol{\omega}$, provides

$$\boldsymbol{\omega} = \nabla\times\boldsymbol{v} = \begin{cases} \begin{pmatrix} \dfrac{\partial v_3}{\partial x_2} - \dfrac{\partial v_2}{\partial x_3} \\ \dfrac{\partial v_1}{\partial x_3} - \dfrac{\partial v_3}{\partial x_1} \\ \dfrac{\partial v_2}{\partial x_1} - \dfrac{\partial v_1}{\partial x_2} \end{pmatrix} & \text{3D Cartesian} \\ \begin{pmatrix} 0 \\ 0 \\ \dfrac{\partial v_2}{\partial x_1} - \dfrac{\partial v_1}{\partial x_2} \end{pmatrix} & \text{2D Cartesian} \\ \mathbf{0} & \text{1D Cartesian} \\ \begin{pmatrix} \dfrac{1}{r}\dfrac{\partial v_z}{\partial\phi} - \dfrac{\partial v_\phi}{\partial z} \\ \dfrac{\partial v_r}{\partial z} - \dfrac{\partial v_z}{\partial r} \\ \dfrac{1}{r}\dfrac{\partial}{\partial r}(r\,v_\phi) - \dfrac{1}{r}\dfrac{\partial v_r}{\partial\phi} \end{pmatrix} & \text{cylindrical } (r,\phi,z) \end{cases} \tag{2.75}$$

which represents the rotation of the vector field $\boldsymbol{v}$. Gradient, divergence and curl, respectively, are sometimes written in other notations, such as

$$\nabla = \mathbf{grad} \qquad \nabla\cdot = \mathrm{div} \qquad \nabla\times = \mathbf{curl} \tag{2.76}$$

Axisymmetric problems written in cylindrical coordinates represent a specific case. Axisymmetry assumes that all flow components along the azimuthal direction ϕ vanish whereby the domain of interest Ω is reduced to a 2D meridional domain in (r, z). Under such conditions the above relationships (2.71)–(2.75) can be significantly simplified with $v_\phi = 0$, $\frac{\partial}{\partial\phi} = 0$.

2.1.7 Gauss's Integral Theorem

The Gauss's integral (or divergence) theorem represents the most valuable transformation in tensor analysis. It relates volume integral to surface integral expressions. Let Ω be the volume of a domain which is bounded by a piecewise-smooth closed surface Γ, let $\boldsymbol{n}\,(\equiv\, \hat{\boldsymbol{n}})$ be the outward-directed unit normal to Γ (Fig. 2.8) and assuming the (scalar) variable ψ and the vector field $\boldsymbol{a}$ have continuous first partial derivatives in Ω, then

$$\begin{aligned}
\int_\Omega \nabla\psi\, d\Omega &= \int_\Gamma \psi\, \boldsymbol{n}\, d\Gamma \\
\int_\Omega \nabla\cdot\boldsymbol{a}\, d\Omega &= \int_\Gamma \boldsymbol{a}\cdot\boldsymbol{n}\, d\Gamma \\
\int_\Omega \nabla\cdot(\boldsymbol{a}\psi)\, d\Omega &= \int_\Gamma \psi(\boldsymbol{a}\cdot\boldsymbol{n})\, d\Gamma \\
\int_\Omega \nabla\times\boldsymbol{a}\, d\Omega &= \int_\Gamma \boldsymbol{a}\times\boldsymbol{n}\, d\Gamma \\
\int_\Omega \nabla\times(\boldsymbol{a}\psi)\, d\Omega &= \int_\Gamma \psi(\boldsymbol{a}\times\boldsymbol{n})\, d\Gamma
\end{aligned} \tag{2.77}$$

where $\int_\Omega (.) d\Omega$ represents a volume integral and $\int_\Gamma (.) d\Gamma$ a surface integral. Using partial integration

$$\nabla\cdot(\boldsymbol{a}\,\psi) = \boldsymbol{a}\cdot\nabla\psi + \psi(\nabla\cdot\boldsymbol{a}) \tag{2.78}$$

we obtain with (2.77)

$$\int_\Omega \psi(\nabla\cdot\boldsymbol{a})\, d\Omega = -\int_\Omega (\nabla\psi\cdot\boldsymbol{a})\, d\Omega + \int_\Gamma \psi(\boldsymbol{a}\cdot\boldsymbol{n})\, d\Gamma \tag{2.79}$$

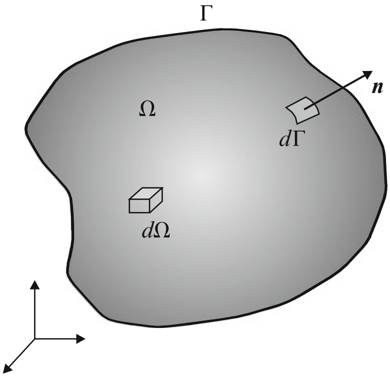


Fig. 2.8 Domain of volume Ω with its closed boundary of surface Γ and the outward unit normal $\boldsymbol{n}$ to the surface

2.1.8 Stokes' Theorem

The Stokes' theorem relates a surface integral Γ over a cap to a line integral S around a bounding curve. The theorem states that the total circulation $\boldsymbol{a} \cdot \boldsymbol{t}$ of a vector field $\boldsymbol{a}$ in form of the line integral is equal to the surface integral of the normal component of $\nabla \times \boldsymbol{a}$:

$$\int_{\Gamma} (\nabla \times \boldsymbol{a}) \cdot \boldsymbol{n} \, d\Gamma = \int_{S} \boldsymbol{a} \cdot \boldsymbol{t} \, dS \tag{2.80}$$

where $\boldsymbol{t}$ is the unit tangent vector.

2.1.9 Reynolds' Transport Theorem

The Reynolds' transport theorem (as a generalization of the Leipniz's integral rule) is very useful to compute derivatives of integrated quantities such as

$$\mathcal{F}(t) = \int_{\Omega} f(\boldsymbol{x}, t) \, d\Omega \tag{2.81}$$

Let $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$ be a fluid vector field and let $\Omega = \Omega(t)$ be a volume bounded by a closed surface $\Gamma = \Gamma(t)$ moving with the fluid, then

$$\frac{D}{Dt} \int_{\Omega(t)} f(\boldsymbol{x}, t) \, d\Omega = \int_{\Omega(t)} \frac{\partial f}{\partial t} \, d\Omega + \int_{\Gamma(t)} f \, (\boldsymbol{v} \cdot \boldsymbol{n}) \, d\Gamma \tag{2.82}$$

or with (2.77)

$$\frac{D}{Dt}\int_{\Omega(t)} f(\boldsymbol{x},t)\,d\Omega = \int_{\Omega(t)}\left[\frac{\partial f}{\partial t} + \nabla\cdot(\boldsymbol{v}\,f)\right]d\Omega \tag{2.83}$$

where $\frac{D}{Dt}$ is the material derivative (2.38) and $\boldsymbol{n}$ is the outward pointing normal vector on the surface Γ (Fig. 2.8).

2.1.10 Classification of Vector Fields

A vector field $\boldsymbol{v}$ is called *solenoidal*, or divergenceless, if

$$\nabla\cdot\boldsymbol{v} = 0 \tag{2.84}$$

By Gauss's theorem (2.77) this is equivalent to

$$\int_\Gamma \boldsymbol{v}\cdot\boldsymbol{n}\,d\Gamma = 0 \tag{2.85}$$

for any closed surface Γ.

A vector field $\boldsymbol{v}$ is called *irrotational*, or cureless, if

$$\boldsymbol{\omega} = \nabla\times\boldsymbol{v} = \boldsymbol{0} \tag{2.86}$$

By Stokes' theorem (2.80) this is equivalent to

$$\int_S \boldsymbol{v}\cdot\boldsymbol{t}\,dS = 0 \tag{2.87}$$

for every closed curve S. A flow satisfying (2.86) is called *potential flow*.

2.1.11 Potential Function, Streamfunction, Streamline and Pathline

It is possible to construct scalar functions on which either a solenoidal (2.84) or an irrotational (2.86) vector field can be implicitly satisfied. Assuming the vector field $\boldsymbol{v}$ could be the gradient of a (scalar) *potential function* Φ in a form such as

$$\boldsymbol{v} = -\nabla\Phi, \tag{2.88}$$

it is easy to show that (2.88) satisfies irrotationality (2.86), i.e., $\boldsymbol{\omega} = \mathbf{0}$, because $-\nabla \times \nabla \Phi = \mathbf{0}$. This is true for all dimensions, cf. (2.75). Introducing (2.88) the potential flow holds with $\nabla \cdot \boldsymbol{v}$:

$$-\nabla^2 \Phi = 0 \tag{2.89}$$

We note that an alternate gradient expression such as $\boldsymbol{v} = -K(\boldsymbol{x})\nabla h$ does not strictly satisfy $\boldsymbol{\omega} = \mathbf{0}$, where $K(\boldsymbol{x})$ is a spatially dependent coefficient and h could be a (different) scalar function, which can be seen as a *pseudopotential* function. If (and only if) $K = \text{const}$, and hence $\Phi = K\,h$, the flow is irrotational (see [33] for more discussions).

On the other hand, we can find a function Ψ, called as *streamfunction*, written in 2D and axisymmetric flow as

$$\begin{aligned} v_1 &= \frac{\partial \Psi}{\partial x_2}, \quad v_2 = -\frac{\partial \Psi}{\partial x_1} \qquad \text{2D Cartesian} \\ v_r &= \frac{1}{r}\frac{\partial \Psi}{\partial z}, \quad v_z = -\frac{1}{r}\frac{\partial \Psi}{\partial r} \qquad \text{axisymmetric} \end{aligned} \tag{2.90}$$

which implicitly satisfies the condition of a selenoidal vector field $\nabla \cdot \boldsymbol{v} = 0$ (2.84) applied to 2D and axisymmetric problems. Inserting (2.90) into the vorticity equation (2.75) the Laplacian equation holds

$$-\nabla^2 \Psi = 0 \quad \text{2D and axisymmetric} \tag{2.91}$$

to determine the streamfunction Ψ for 2D and axisymmetric flows. For 3D problems it is not possible to find a scalar function capable of satisfying a divergenceless velocity, $\nabla \cdot \boldsymbol{v} = 0$ similar to 2D. We emphasize that *a streamfunction analog doesn't exist for 3D problems*.

By definition, a *streamline* in a flow is defined as a line, which, at any instant, is tangent to the velocity vector $\boldsymbol{v}$. If $d\boldsymbol{x}$ is a differential along a streamline, the tangency condition is expressed by the cross product $\boldsymbol{v} \times d\boldsymbol{x} = \mathbf{0}$. In 3D Eulerian coordinates it reads to:

$$\frac{dx_1}{v_1(\boldsymbol{x}, t_a)} = \frac{dx_2}{v_2(\boldsymbol{x}, t_a)} = \frac{dx_3}{v_3(\boldsymbol{x}, t_a)} \tag{2.92}$$

where t_a indicates a certain (fixed) time. The cross product of the two nonzero vectors $\boldsymbol{v}$ and $d\boldsymbol{x}$ is zero only if they are parallel. Accordingly, a unique direction for the streamline exists at all points in space $\boldsymbol{x}$, provided $\boldsymbol{v}$ is not zero. However, an exception is given at so-called stagnation points, where the velocity $\boldsymbol{v}$ is zero. At those points streamlines can be split into two or more streamlines. Once the velocity field $\boldsymbol{v}$ is known the solution of (2.92) yields a family of streamlines, referred to as the motion pattern. Only in 2D (or axisymmetry) a streamline can be identified as a graph of constant values of streamfunction $\Psi = \Psi(\boldsymbol{x})$ (Fig. 2.9).

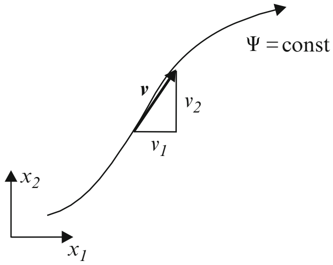


Fig. 2.9 The streamfunction Ψ as a streamline in 2D flow

The streamfunction in 2D must obey the general differential relation:

$$d\Psi = \frac{\partial \Psi}{\partial x_1} dx_1 + \frac{\partial \Psi}{\partial x_2} dx_2 \tag{2.93}$$

Substituting (2.90) into (2.93) gives

$$d\Psi = -v_2\, dx_1 + v_1\, dx_2 \tag{2.94}$$

Accordingly, for a streamline with constant Ψ, i.e. $d\Psi = 0$, (2.94) becomes

$$\frac{v_2}{v_1} = \left.\frac{dx_2}{dx_1}\right|_{\Psi=\text{const}} \tag{2.95}$$

showing that the velocity vector is tangent (2.92) to the curve $\Psi = \text{const}$.

A *pathline* is a curve (or line) along which a fixed massless particle moves during a sequence of times t. It is thus the *trajectory* of a particle of fixed identity. In the Eulerian formulation the differential equation for a pathline directly results from (2.39):

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{v}(\boldsymbol{x}, t) \tag{2.96}$$

or written in 3D Eulerian coordinates

$$\frac{dx_1}{v_1(\boldsymbol{x}, t)} = \frac{dx_2}{v_2(\boldsymbol{x}, t)} = \frac{dx_3}{v_3(\boldsymbol{x}, t)} = dt \tag{2.97}$$

The solution of (2.96), or (2.97), for a particle location at any time t can be expressed by

$$\boldsymbol{x}(t) = \boldsymbol{x}(t_0) + \int_{t_0}^{t} \boldsymbol{v}\big(\boldsymbol{x}(t), t\big)\, dt \tag{2.98}$$

where $\boldsymbol{x}(t_0)$ is the position at initial time t_0.

Finally, we note that under transient flow conditions, i.e., where the flow field is time-dependent, streamlines and pathlines are commonly distinct. Here we can see an instantaneous picture of the streamlines, as the picture varies continuously. However, for steady-state flow conditions, i.e., where flow characteristics remain invariant with time, streamlines and pathlines coincide.

2.2 Classifications and Definitions

Flow and transport processes in the context of subsurface modeling are usually related to terms and descriptions which will be summarized in the following. The most important definitions are presented which we will need to relate to in the subsequent chapters. For a more comprehensive discussion of basic definitions for porous-media and groundwater problems the reader is referred to Bear [34] or Bear and Cheng [38].

2.2.1 Water and Aquifer

Subsurface water denotes *all* water below the ground surface (Fig. 2.10). This water is contained in the void space of geologic formations of different types. The void space can be fully or partially saturated by water. Subsurface water can be regarded as part of the hydrologic cycle [356].

Groundwater denotes only this part of the subsurface water that occurs in geologic formations in which the void space is *fully saturated*. Groundwater flows in aquifers and rocks.

Surface water denotes all water collecting on the ground or in streams, rivers, lakes, wetlands or oceans. Surface water is usually interrelated to subsurface water where water can be exchanged via infiltration, drainage and seepage.

Freshwater represents surface or subsurface water having only low concentrations of salts or other dissolved solids. Commonly, a groundwater resource (without additional specifications) is related to freshwater which is available for drinking and other purposes. Freshwater specifically excludes saltwater. Measuring water salinity by concentrations in parts per million (ppm) – equivalent to g/l – freshwater is usually classified with a concentration smaller than 0.5 ppm.

Saltwater is water which contains dissolved salts (mostly NaCl) of different concentrations larger than 0.5 ppm. It can be further categorized into *brackish water* having a salinity in the range of 0.5–30 ppm, *saline water* with a salinity between 30 and 50 ppm and *brine* with a salinity of more than 50 ppm.

Saltwater intrusion (or saltwater encroachment) denotes the movement of saltwater into freshwater. In the subsurface it virtually occurs in all coastal aquifers, where the denser saltwater from the sea intrudes into the freshwater aquifer due to

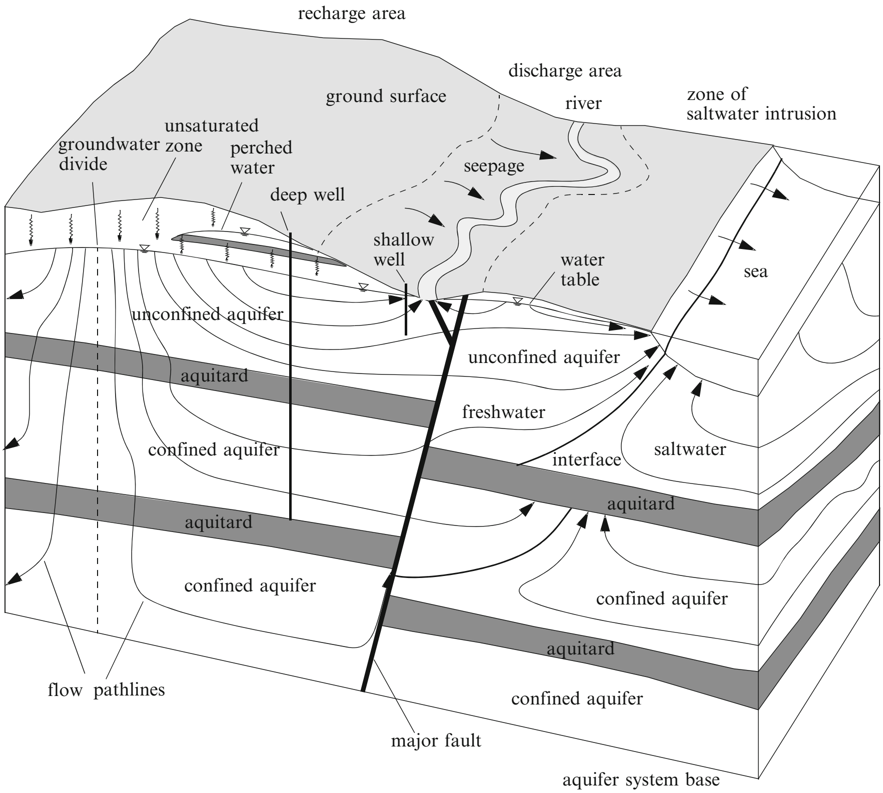


Fig. 2.10 Illustration of typical flow regimes in a multilayered aquifer system (Modified from [101, 306])

its higher density. It can also be caused by groundwater pumping above or nearby saltwater zones.

Aquifer denotes a geologic formation, or a group of formations, of water-bearing permeable rock or sediment layers from which water can be usefully extracted (Fig. 2.10). Aquifers can be confined or unconfined (see further below).

Aquitard is a geologic formation which is of a semipervious nature. It transmits water at a very low rate compared to an aquifer. An aquitard separates an aquifer layer from an adjacent aquifer (as exemplified in Fig. 2.10). An aquitard, if completely impermeable, is denoted as *aquiclude* or *aquifuge*.

Aquifer system groups a certain number of aquifers separated by aquitards in a multilayered structure (Fig. 2.10).

Confined aquifer, also known as *pressure aquifer*, is an aquifer (a) bounded from above and below by impervious formations and an aquifer (b) in which the water pressure reaches such values that the water level measured in a piezometer will rise above the base of the upper confining formation. Water enters a confined aquifer in a recharge area, which is commonly linked to an unconfined aquifer. A confined aquifer is called a *leaky confined aquifer* if one or both confining formations are semipervious, through which leakage may take place.

Unconfined aquifer, also called a *phreatic aquifer*, is bounded from above by the water table or phreatic surface. Usually, a phreatic aquifer is directly recharged from the ground surface above it, except where impervious layers (of limited areal extent) exist between the phreatic surface and the ground surface. Above the phreatic surface a capillary fringe establishes. The base of an unconfined aquifer is considered impervious. An unconfined aquifer is called a *leaky unconfined aquifer* if the lower bounded formation is semipervious.

Perched groundwater, or *perched aquifer*, is a special case of a phreatic aquifer. It represents a limited areal extent of water, formed on an impervious, or semipervious, layer (see Fig. 2.10). Perched water may exist only for a limited period of time.

Saturated zone forms above impervious or semipervious formations. In this zone the entire void space is filled with water. The saturated zone can be bounded from above by a water table, or phreatic surface.

Unsaturated zone, or *vadose zone*, describes the zone between ground surface and the underlying phreatic surface, where only part of the void space is occupied by water, the remainder being occupied by a gaseous phase, usually air.

Infiltration is the unsaturated downward water flow from the ground surface, percolating through the unsaturated zone and reaching an underlying water table. It is usually driven by natural replenishment from precipitation and snow melt. Its quantity in relation to the total precipitation is influenced, among others, by evaporation, surface runoff, soil characteristics and transpiration through the vegetation. Infiltration can also include seepage from ponds, lakes, ditches, channels and other leakages.

Groundwater recharge denotes that amount of infiltrating water which finally reaches the water table of an underlying aquifer. It determines the replenishment of aquifers and represents an important parameter in the use and exploitation of groundwater resources.

Groundwater divide is a surface in 3D or a curve in 2D that separates the flow domain into subdomains, on either side of which groundwater moves in opposite directions (see Fig. 2.10).

Water table, or *phreatic surface*, is actually the boundary between the unsaturated and saturated zone. It represents the upper surface of the groundwater body. Phreatic surface is a specific representation of a *free surface*.

Fracture is part of the void space of a porous-medium domain that has a special spatial configuration: one of its dimensions – the *aperture* – is much smaller than the other two spatial dimensions. Fractures provide pathways for fluid flow and transport through otherwise impermeable or semipervious formations and produce planes, surfaces or even lined interconnections where fluid movement increases and focuses, such as in cracks of rocks, interstices, vugs or tectonic faults.

Fractured porous rock defines a pervious rock formation which is composed of an interconnected network of fractures. Thus, the total void space results from fractures and porous blocks of rock. The flow movement usually dominates in the fracture network. If the surrounded rock contains no void space the term *fractured rock* is used.

2.2.2 Terms and Quantities

Domain and **boundary** are defined in the $D-$dimensional Euclidean space $\Re^D$ ($D = 1, 2, 3$) (see Sect. 2.1.2) and are usually denoted by $\Omega \subset \Re^D$ and $\Gamma \subset \Re^D$, respectively. The domain Ω is completely closed by the boundary Γ (see Fig. 2.8). The boundary Γ can be arbitrarily shaped. It can be composed of different nonoverlapping segments, e.g., $\Gamma_D, \Gamma_N, \ldots$, bounding the domain Ω both outside and inside. By definition, the boundary Γ is separated from the domain Ω. On the other hand, by $\bar{\Omega}$ we denote the (closure) domain, which completely joins the boundary

$$\bar{\Omega} = \Omega \cup \Gamma \quad \text{with} \quad \Gamma = \Gamma_D \cup \Gamma_N \cup \ldots \tag{2.99}$$

On $\bar{\Omega}$ and Γ initial conditions and boundary conditions have to be specified, respectively.

Initial conditions (IC's) specify the values of a time-dependent variable $\psi = \psi(\boldsymbol{x}, t)$ in $\bar{\Omega}$ at initial time t_0:

$$\psi(\boldsymbol{x}, t_0) = \psi_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\Omega} = \Omega \cup \Gamma \tag{2.100}$$

We note that for steady-state problems no IC's are needed, unless the steady-state problem is nonlinear, where IC's are required to initialize an iterative procedure.

Boundary conditions (BC's) specify values on the total boundary Γ closing the domain Ω. Related to a time-dependent solution variable $\psi = \psi(\boldsymbol{x}, t)$ the following types are common:

1. *Dirichlet-type* (1st kind) BC, also termed as *essential boundary condition*, prescribes the value of ψ on a boundary section Γ_D:

$$\psi(\boldsymbol{x}, t) = \psi_D(t) \quad \text{on} \quad \Gamma_D \subset \Gamma \tag{2.101}$$

   A typical application of a Dirichlet BC is the prescription of a potential value, mass concentration or temperature in dependence on the underlying problem.

2. *Neumann-type* (2nd kind) BC prescribes the normal derivative of ψ on a boundary section Γ_N:

$$-(\boldsymbol{\alpha} \cdot \nabla \psi) \cdot \boldsymbol{n} = q_n(t) \quad \text{on} \quad \Gamma_N \subset \Gamma \tag{2.102}$$

where $\boldsymbol{\alpha}$ is an arbitrary coefficient matrix, which must be $\|\boldsymbol{\alpha}\| \neq 0$. The prescribed value $q_n(t)$ represents a normal flux (positive outward) across the boundary portion Γ_N. If $q_n = 0$ the Neumann-type BC reduces to a *natural* (no-flux) boundary condition associated with $\nabla\psi = \mathbf{0}$. A typical application of a Neumann BC is the description of a diffusive (dispersive/conductive) flux rate of mass or energy in dependence on the underlying problem.

3. *Cauchy-type* and *Robin-type* (3rd kind) BC's combine Dirichlet-type and Neumann-type BC's in different ways. The Cauchy BC represents a weighted arithmetic mean of Dirichlet and Neumann BC according to

$$-(\boldsymbol{\alpha} \cdot \nabla\psi) \cdot \boldsymbol{n} = -\beta(\psi_C - \psi) \quad \text{on} \quad \Gamma_C \subset \Gamma \tag{2.103}$$

where $\psi_C = \psi_C(t)$ is a prescribed value of ψ and $\beta = \beta(t)$ denotes an additional transfer coefficient. We have chosen the signs in (2.103) in such a way that the flux is directed positive outward if $\psi > \psi_C$. If β becomes large the boundary condition tends to a Dirichlet type with $\psi \to \psi_C$ on Γ_C. On the other hand, if β becomes small it tends to a natural boundary condition enforcing $\nabla\psi \to \mathbf{0}$ on Γ_C. A typical application for Cauchy BC is the leakage of mass through a given boundary section. In contrast, the Robin-type (also called *mixed*) BC is a mixture of Dirichlet and Neumann BC in such a form

$$-(\boldsymbol{\alpha} \cdot \nabla\psi) \cdot \boldsymbol{n} = q_n - \beta(\psi_C - \psi) \quad \text{on} \quad \Gamma_R \subset \Gamma \tag{2.104}$$

The Robin-type BC is the most general boundary condition and implies all other types of boundary conditions above. A typical application of Robin BC refers to the prescription of a total (diffusive plus advective) mass or energy rate through a given boundary section. If the total (diffusive plus advective) mass or energy rate q_n is specified in such a form

$$-(\boldsymbol{\alpha} \cdot \nabla\psi - \boldsymbol{v}\psi) \cdot \boldsymbol{n} = q_n \tag{2.105}$$

where $\boldsymbol{v} \cdot \boldsymbol{n}$ is the advective velocity normal (outward positive) on Γ_C, (2.105) can be expressed by a Cauchy-type BC (2.103) if we substitute

$$\beta = -\boldsymbol{v} \cdot \boldsymbol{n} \quad \text{and} \quad \psi_C = \frac{q_n}{\boldsymbol{v} \cdot \boldsymbol{n}} \tag{2.106}$$

We note that the union of Γ_D, Γ_N, Γ_C and Γ_R forms the complete boundary $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_C \cup \Gamma_R$, where the segments do not overlap each other, $\Gamma_D \cap \Gamma_N \cap \Gamma_C \cap \Gamma_R = \emptyset$. Boundary conditions are always required for both transient and steady-state problems. Usually, $\Gamma_D \neq \emptyset$ at steady-state. Besides these boundary conditions of Dirichlet (2.101), Neumann (2.102), Cauchy (2.103) or Robin (2.104) type, there are more specific boundary conditions, for example *free surface*, *seepage face*, *surface ponding*, *pumping well*, *borehole heat exchanger* or *gradient-type* BC, which represent modifications and in parts nonlinear combinations of the above conditions and will be described in the context of the problem solutions to be discussed in Chap. 6.

Transfer or **leakage** conditions describe the exchange of mass, momentum or energy on specific boundaries. Transfer is a more general term typically used in mass and heat transport, while leakage is commonly used in subsurface hydrology to describe the exchange of flow through external and internal boundaries. Generally expressed, their mathematical formulation reads to

$$q_n = -\beta(\psi_{\text{ex}} - \psi) \tag{2.107}$$

written for the variable ψ, where β represents a transfer coefficient. Accordingly, the normal exchange rate q_n is controlled by the difference between an external (known) value ψ_{ex} and the internal value of the variable ψ and can be recognized as a Cauchy-type BC, cf. (2.103), with $q_n = -(\boldsymbol{\alpha} \cdot \nabla\psi) \cdot \boldsymbol{n}$. Applied to the heat transfer, Equation (2.107) is known as *Newton's law of cooling*, where heat transfer occurs at a boundary of a solid with the ambient convecting fluid temperature. Equation (2.107) also represents a leakage condition, where the transfer coefficient is usually replaced by

$$\beta = \frac{K}{b} \tag{2.108}$$

introducing the conductivity K and the thickness b. In practice, the transfer coefficient β may be chosen to distinguish between inflowing conditions ($q_n < 0$) and outflowing conditions ($q_n > 0$) introducing the following *dual* functions for the transfer coefficient β:

$$\beta = \begin{cases} \beta_{\text{in}} & \text{for} \quad \psi_{\text{ex}} > \psi \\ \beta_{\text{out}} & \text{for} \quad \psi_{\text{ex}} \leq \psi \end{cases} \tag{2.109}$$

where β_{in} and β_{out} denote the in-transfer and the out-transfer coefficients, respectively.

Resistance of mass, momentum or heat exchange at a surface is related to the inverse of the transfer coefficient (2.108)

$$\bar{R} = \frac{1}{A\beta} \tag{2.110}$$

where A corresponds to an exchange area.

Specific resistance defines the resistance of mass, momentum or heat exchange per unit length such that

$$R = \bar{R}\,L = \frac{L}{A\beta} = \frac{1}{S\beta} \tag{2.111}$$

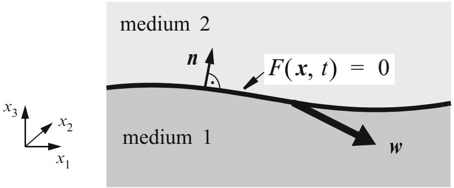


Fig. 2.11 Interface representation

where L is a length and $S = \frac{A}{L}$ is a specific exchange surface. The specific resistance represents a material property.

Interface, or *interface boundary* or *surface*, represents a boundary between two media (materials) where the conditions abruptly change (Fig. 2.11). This can be a boundary between different fluids or a boundary between a fluid and a solid. If liquids are immiscible, a distinct sharp interface is maintained between them, even if small quantities of certain components can cross the interphase boundary driven by diffusion. Mathematically, an interface F can be described by the function:

$$F(\boldsymbol{x}, t) = 0 \tag{2.112}$$

As the interface moves with a velocity $\boldsymbol{w}$, its shape changes, however, all material points associated with the interface must be conserved, i.e., the material derivative of F is valid

$$\frac{\partial F}{\partial t} + \boldsymbol{w} \cdot \nabla F = 0 \tag{2.113}$$

The outward unit vector $\boldsymbol{n}$ normal to F is defined as

$$\boldsymbol{n} = \frac{\nabla F}{\|\nabla F\|} \tag{2.114}$$

and accordingly it is

$$\boldsymbol{w} \cdot \boldsymbol{n} = -\frac{\partial F / \partial t}{\|\nabla F\|} \tag{2.115}$$

Free surface is a specific surface of a connected flow domain that is subjected to a constant pressure and a given mass flux crossing the surface. Having a zero mass flux it represents the classic hydrodynamic free-surface condition for an isobaric and impervious boundary. In subsurface flow a phreatic surface represents a free surface.

Phase, identified by the index α (or other Greek indices), is defined as a portion Ω^α of space Ω, whether connected or non-interconnected, that is separated from other such portions by a well defined surface Γ^α, which represents an interface. A phase

α may be composed of a number of different *chemical species* k. The phase index α takes on values of s and $f = (l, g)$ corresponding to the solid phase and the two fluid phases of liquid and gas, respectively. Throughout this book, α (or other Greek indices) ranges as $\alpha = s, f \in (l, g)$ and the repetition of Greek indices does not imply a summation.

Energy in physics represents a quantity that is assigned to a particle, an object and a system of objects. Energy is a scalar physical quantity, which is usually measured in joules (J). There are different forms of energy, e.g., internal or thermal energy and kinetic energy, which are named after the related forces. All forms of energy are equivalent. Energy is subject to a conservation law, the *first law of thermodynamics*. Any form of energy can be transformed into another form, however, in the energy transformation process the total energy remains the same. Energy may neither be created nor destroyed.

Entropy is a measure of how disorganized a system is. It is taken as a measure of 'disorder': the higher the entropy, the higher the disorder. Disorder means the tendency of a system to get states which are homogeneous and fully mixed throughout space. The highest degree of disorder is the chaos, the most unorganized state. Entropy of a physical system is proportional to the quantity of energy no longer available to do physical work. It is measured in physical units of energy per temperature: joules per kelvin (J/K). Entropy is central to the *second law of thermodynamics*, which states that in an isolated system any activity increases the entropy. The second law of thermodynamics introduces irreversibility: an isolated system cannot pass from a state of higher entropy to a state of lower entropy, e.g., transmission of heat from a cooler medium to a warmer one is impossible. Increases in entropy correspond to irreversible changes in a system. Entropy reaches its maximum at equilibrium state of a physical system.

Chemical species, or a *component*, identified by subscript k, is a part of a phase that consists in an identifiable, chemical constituent, or an assembly of constituents, e.g., ions or molecules. It represents a mixture of a number of independent chemical species ($k = 1, \ldots, N$) dissolved in a fluid phase or adsorbed at/absorbed in a solid phase. Chemical species are miscible continuous quantities, which cannot be separated by interface (discontinuity) conditions. Note that Einstein's summation convention is not applied to the species index k.

Extensive quantity, $\mathcal{F}_k(t)$, specified for chemical species k, such as for mass $\mathcal{M}_k$, momentum $\mathcal{V}_k$, internal (thermal) energy $\mathcal{E}_k$, kinetic energy $\mathcal{K}_k$ and entropy $\mathcal{S}_k$, is given for the domain Ω and reads

$$\mathcal{F}_k(t) = \int_{\Omega} f_k(\boldsymbol{x}, t)\, d\Omega \tag{2.116}$$

where $f_k(\boldsymbol{x}, t)$ is the *intensive quantity* of species k. Extensive quantities are dependent on the volume Ω and listed in Table 2.1.

Intensive quantity, $f_k(\boldsymbol{x}, t)$ as listed in Table 2.1, is related to the particles occupied in the domain Ω. It represents the intensity of an extensive quantity as

Table 2.1 Extensive $\mathcal{F}_k$ and intensive f_k quantities related to species k (no summation over k)

Quantity	$\mathcal{F}_k(t)$	$f_k(\boldsymbol{x}, t)$
Mass	$\mathcal{M}_k$	ρ_k
Momentum	$\mathcal{V}_k$	$\rho_k\, \boldsymbol{v}_k$
Internal energy	$\mathcal{E}_k$	$\rho_k\, E_k$
Kinetic energy	$\mathcal{K}_k$	$\rho_k\, \frac{1}{2} v_k^2$
Entropy	$\mathcal{S}_k$	$\rho_k\, S_k$

defined by (2.116) and is accordingly independent of the volume Ω. In general an intensive quantity that is given per unit volume, is characterized as a *density*. On the other hand, an intensive quantity as given per unit mass, will be denoted as a *specific density*. In Table 2.1 it can be recognized that the velocity $\boldsymbol{v}_k$ represents a specific momentum density, E_k is a specific internal energy density and S_k is a specific entropy density.

Density and **specific density** represent intensive quantities per unit volume and unit mass, respectively. *Mass density*, ρ_k, is defined as mass per unit volume, *specific momentum density*, $\boldsymbol{v}_k$, is momentum per unit mass, *specific internal energy density*, E_k, denotes internal energy per unit mass and *specific entropy density*, S_k, is entropy per unit mass. Commonly, for short descriptions it is customary to use the word *density* also when we actually mean mass density ρ_k.

Concentration measures the quantity of chemical species k in a unit volume Ω of fluid. It can be expressed in different ways as follows.

Mass concentration, denoted by C_k, expresses the mass of species k per unit volume of a fluid and is identical to the mass density ρ_k:

$$C_k \equiv \rho_k = \frac{\mathcal{M}_k}{dV} \tag{2.117}$$

where dV corresponds to an averaging volume. This measure is preferably used. Usual physical units are g/l (= grams of k per liter of fluid), or mg/l (= milligrams of k per liter of fluid).

Molar concentration, or *molarity*, denoted by $[C_k]$, expresses the number of $k-$moles per unit volume of fluid and reads

$$[C_k] = \frac{C_k}{m_k} \tag{2.118}$$

where m_k is the molecular mass of the $k-$species. This measure is common for thermodynamics. Common units are moles of k per liter of fluid, mol/l or mol/m^3($\equiv$ mmol/l).

Activity of a species k, denoted by $\{C_k\}$, is related to its molar concentration $[C_k]$ by

$$\{C_k\} = \gamma_k [C_k] \tag{2.119}$$

where γ_k is the *activity coefficient* of species k, which is given for an ionic aqueous species for instance by the empirical *Davies relationship* [500] in the form

$$\log_{10} \gamma_k = -\tfrac{1}{2} z_k^2 \Big(\frac{\sqrt{I}}{1 + \sqrt{I}} - 0.3\, I \Big) \tag{2.120}$$

where z_k is the charge on the kth species and I denotes the *ionic strength* defined by

$$I = \tfrac{1}{2} \sum_{k=1}^{N} z_k^2 \; [C_k] \tag{2.121}$$

For *dilute solutions*, $\gamma_k \approx 1$ and

$$\{C_k\} \approx [C_k] \tag{2.122}$$

Mass fraction, denoted by ω_k, is the mass of $k-$species per unit mass of fluid. It can be seen as a *specific density* of mass and is expressed as

$$\omega_k = \frac{C_k}{\rho} \equiv \frac{\rho_k}{\rho} \qquad \rho = \sum_k \rho_k \qquad \sum_k \omega_k = 1 \tag{2.123}$$

where ρ is the bulk mass density of fluid. This dimensionless measure is often expressed with the physical unit ppm, ‘parts per million’, defining the number of grams of solute per million grams of solution.

Advection describes the transport mechanism of a conserved quantity (e.g., mass or heat) due to fluid motion. Advection requires currents in the fluid (or fluid phase). It does not occur in impervious media or stagnant fluids.

Convection is sometimes synonymously used with advection. However, it usually refers to more general flow phenomena, in which the fluid motion is additionally influenced or even caused by changes in the fluid (mass) density. One can differentiate between *forced convection* in which the fluid motion is generated by external forces (e.g., pressure/potential gradient, flow source), *free (or natural) convection* in which the flow motion exclusively results from inner buoyant forces due to fluid density changes and *mixed convection* where forced and free convection occur in combination. Throughout the book we will use the term ‘convection’ to indicate transport of quantities within a moving fluid, where variable-density effects are present, while the term ‘advection’ is used to indicate transport of quantities in fluid flow without variable-density effects. With this definition we understand that forced convection is equivalent to advection.

Diffusion usually describes the spread of chemical species from regions of higher concentration to regions of lower concentration. It occurs both in fluids and solids. Diffusion in a flowing fluid is independent of the flow direction, i.e., it also acts in the opposite flow direction. More general, diffusion can be understood as a spreading mechanism driven by gradients of one (or even more) quantity(ies). For instance, thermal diffusion is driven by a temperature gradient, mass diffusion is driven by a concentration gradient, momentum diffusion is driven by a gradient of velocity.

Conduction describes a spreading mechanism due to a gradient of a quantity. It is equivalent to diffusion. A hydraulic conduction is controlled by the gradient of a hydraulic potential. A thermal conduction is driven by a temperature gradient.

Sorption is a general term which covers both *adsorption* and *absorption*. Adsorption refers to the adherence of chemical species primarily on a solid surface due to adhesion, while absorption refers to a more or less uniform penetration of chemical species into a coexisting phase. Additionally, there is *desorption* which is the reverse of adsorption, i.e., chemical species are detached from the solid surface and reenter a dissolved phase.

Steady state describes systems, properties or dependent variables which are unchanging in time t. A dependent variable $\psi(\boldsymbol{x}, t)$ becomes *steady* if the time t does not appear as an independent variable anymore, i.e., $\psi = \psi(\boldsymbol{x})$. It implies that its derivative with respect to time is zero:

$$\frac{\partial \psi}{\partial t} = 0 \tag{2.124}$$

Chapter 3
Porous Medium

3.1 Fundamental Concept

The processes of flow, mass and heat refer to extensive quantities (such as mass, momentum, energy and entropy), cf. Sect. 2.2.2, which are transported through a spatial domain of interest. This spatial domain is said to behave as a *continuum* which is occupied by matter for which a continuous distribution can be postulated. The matter may take a number of M aggregate forms or phases α, particularly: solid s, liquid l and gaseous g. It retains their continuity regardless how small volume elements the matter is subdivided in and interior material interfaces or surfaces exist. Any mathematical point we select can be assigned to matter as a physical point of given finite size. In accordance with the assigned size of the physical point we can find, at least, two levels of a continuum description:

1. *Microscopic level*, where every point in the domain is occupied by only one phase (solid or liquid or gaseous).
2. *Macroscopic level*, where properties are defined at every point in the domain consisting of all phases (solid and liquid and gaseous).

At the microscopic level, the basic principles of fluid and solid mechanics can be used to solve the processes in the single phase domain, subject to BC's on the interfaces of phases (e.g., liquid-solid, liquid-gaseous) that bounds this domain. However, at this level the complex interface geometry is neither observable nor describable. Accordingly, the solution of mass and transport processes in porous and fractured media at the microscopic level is impractical and widely impossible to obtain.

At the macroscopic level we can circumvent the difficulties associated with the geometric complexity of coexisting phases, at which measurable and continuous quantities may be determined and BC's can be easily formulated. The continuum approach at such a macroscopic level is obtained via spatial averaging of the phase behaviors over a certain elementary volume. For each point within this macroscopic space, average values for variables and material properties result. The advantage

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_3,

of the macroscopic continuum approach is that (1) there is no need anymore for specifying exact configurations of interphase boundaries, (2) continuous and differentiable quantities result which can be employed by standard mathematical methods, and (3) the macroscopic quantities are measurable and applicable to practical problems.

3.2 Representative Elementary Volume (REV)

The transformation of variables and quantities from the microscopic to the macroscopic level needs spatial averaging referred to a certain elementary volume. It represents an appropriate transition, often termed as *macroscopization*, from a single phase to a multiphase level of description applied to a volume composed of all relevant phases ($\alpha = s, l, g$) (or materials, $M = 3$) of interest. Formally, a porous (or fractured) medium can be defined as a multiphase material body characterized by the following features, e.g., [37]:

1. The averaging volume, denoted by dV, for a porous medium refers to a *representative elementary volume* (REV) which is occupied by a *persistent solid phase s* (Fig. 3.1). The remaining volumetric part, called *void space*, is occupied by one or more fluid phases ($f = l, g$). If such a REV cannot be found for a given domain, that domain cannot qualify as a porous medium.
2. The size of the REV is such that parameters that represent the distribution of the solid phase and void space within it are statistically meaningful.

The porous medium can be naturally formed (e.g., sand beds, rocks, soils) or engineered (e.g., tissues, concretes, polymer composites). Each phase (solid s, liquid l, gaseous g) is regarded as a continuum with smoothly varying properties, overlooking its molecular structure. The REV dV has to be sufficiently large for fluctuations of spatially averaged properties to be negligible. Phases $\alpha = s, f$ are regarded as material subdomains dV_α separated by phase interfaces (e.g., solid-solid, fluid-solid, fluid-fluid). Each phase α is composed of N^α miscible chemical species. It represents a molecular mixture of several identifiable chemical components k. By definition a chemical species k exists in only one phase α. Species that pass through different phases are regarded as separate, phase-pertinent constituents, accordingly the total number of chemical species $N = \sum_\alpha N^\alpha$ holds.

The fundamental assumption of continuum mechanics states that the resulting average quantities have to be independent of the size of the averaging volume dV and have to be continuous over time and space. Thus, the REV region dV is required to possess certain characteristics. Consider, for example, the void space, also termed as *porosity* for a porous medium, $\varepsilon_f = dV_f/dV$. As the size of dV varies, the porosity ε_f varies as shown in Fig. 3.2. If dV is very small, erratic porosity results depending on whether the dV happens to cover voids or solids. Then as dV increases, fluctuations will appear in ε_f because relatively large portions of the one phase or other phases may become part of the averaging domain. As dV further increases

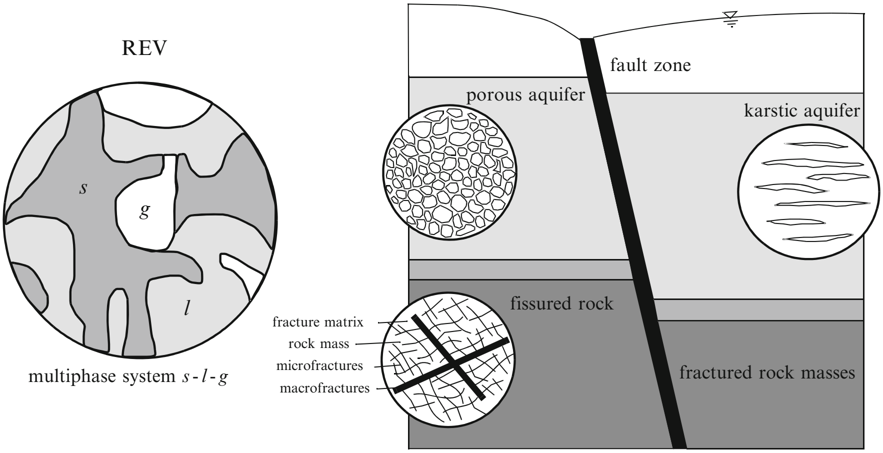


Fig. 3.1 Representation of porous and fractured porous medium

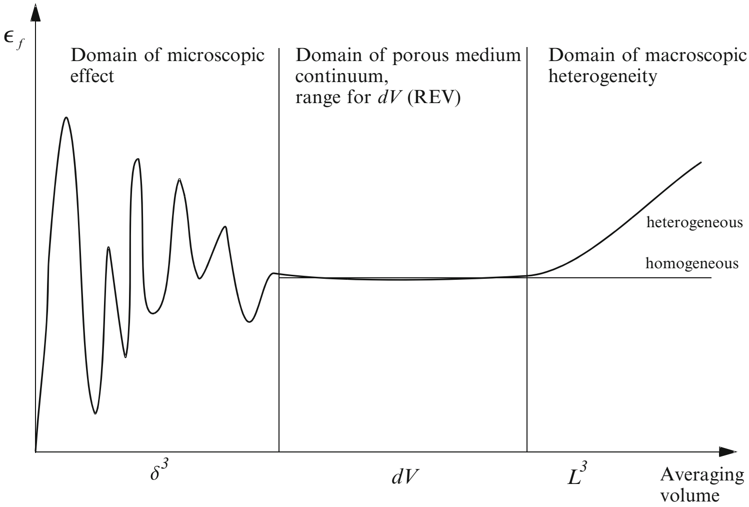


Fig. 3.2 Porosity ε_f as function of averaging volume

within some interval of domain, there is a region when the porosity ε_f remains fairly constant. Within such an interval, in general, all average quantities become independent of the average domain dV. Further increase of dV may cause gross inhomogeneities of the medium that affect the stability of the average (macroscopic) quantities.

In order to maintain meaningful average quantities the characteristic length of the averaging volume dV, denoted by $D \sim dV^{1/3}$, must satisfy the inequality

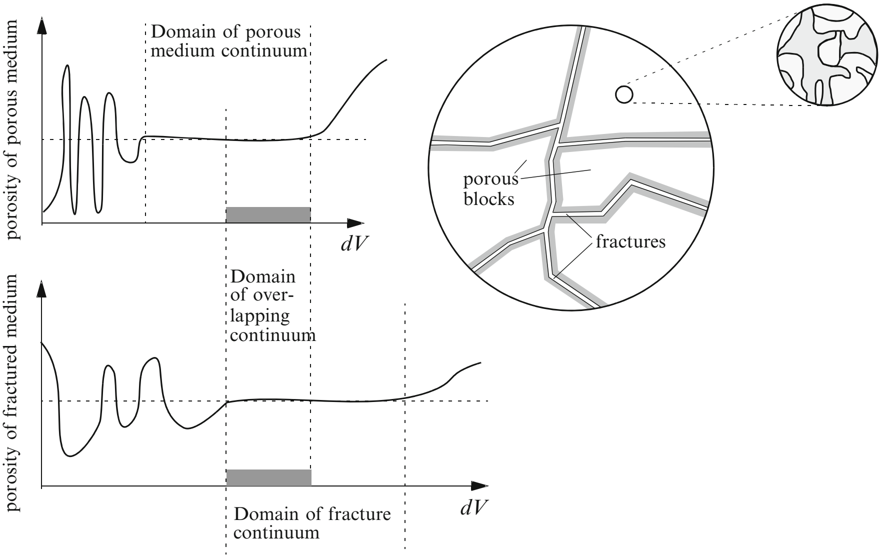

porosity of porous medium
Domain of porous medium continuum
dV
porosity of fractured medium
Domain of over-lapping continuum
Domain of fracture continuum
dV
porous blocks
fractures

Fig. 3.3 REVs for fractured media with overlapping continuum

$$\delta \ll D \ll L \tag{3.1}$$

where δ is the microscopic scale of the medium and L is the scale of the gross inhomogeneities. If for a particular medium these characteristic lengths cannot be identified, if inequality (3.1) does not hold, or if the scale of the problem of interest is of order D, there is no REV and accordingly the averaging technique is not applicable. With other words, the size of a REV must be much larger than the scale of microscopic heterogeneity due to the presence of solid and void space, and much smaller than the scale of the domain of interest (e.g., an aquifer system or a layered domain of soil) having macroscopic heterogeneity.

The features possessing for a REV of a porous medium can also be applied to a REV for the fracture media. In some cases we can recognize that an overlapping REV for the porous media and fractures exists (Fig. 3.3). Then, the problem can be treated as an equivalent continuum. Unlikely, if such an overlapping REV cannot be found (Fig. 3.4), fractures and porous media must be solved in a separate scale and have to be coupled via macroscopic interface conditions by using a discrete fracture approach. This is typical, for example, when fractures' apertures are large, while the voids in the porous blocks are very small, practically all the flow takes place through the fractures. The discrete fracture approach requires information (e.g., aperture, length, orientation, spacing etc.) on every individual fracture. For solving large-scale problems of this type with hundreds or thousands of fractures a huge amount of detailed input data demands and a high computational effort can result.

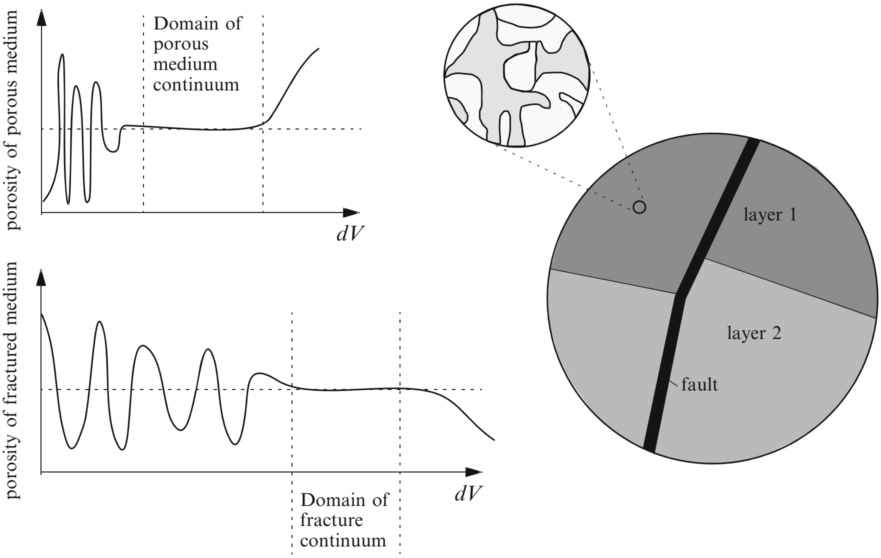

porosity of porous medium
Domain of porous medium continuum
dV
porosity of fractured medium
Domain of fracture continuum
dV
layer 1
layer 2
fault

Fig. 3.4 REVs for fractured media without overlapping continuum

3.3 Average Operators and Average Quantities

The validity of the following averaging procedures is subject to the *existence* of a REV, the averaging volume dV. If valid, we can define an average over dV at *every* mathematical point, denoted by its position vector $\boldsymbol{x}$, independent of whether or not $\boldsymbol{x}$ falls inside the phase. This position vector $\boldsymbol{x}$ serves as the centroid of the REV. On the other hand, let us identify the position of a particle within the REV by the vector $\boldsymbol{r}$ and the position with respect to the centroid of the REV by the vector $\boldsymbol{y}$ as shown in Fig. 3.5:

$$\boldsymbol{r} = \boldsymbol{x} + \boldsymbol{y} \tag{3.2}$$

For each phase a phase distribution function γ_α may be defined as

$$\gamma_\alpha = \gamma_\alpha(\boldsymbol{r}, t) = \begin{cases} 1 & \text{if} \quad \boldsymbol{r} \quad \text{lies in the } \alpha\text{−phase} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \forall t \tag{3.3}$$

The volume volume fraction of the $\alpha-$phase, ε_α, is the fraction of dV occupied by the $\alpha-$phase:

$$\varepsilon_\alpha(\boldsymbol{x}, t) = \frac{dV_\alpha(\boldsymbol{x}, t)}{dV} = \frac{1}{dV} \int_{dV} \gamma_\alpha(\boldsymbol{x} + \boldsymbol{y}, t) dv(\boldsymbol{y}) \tag{3.4}$$

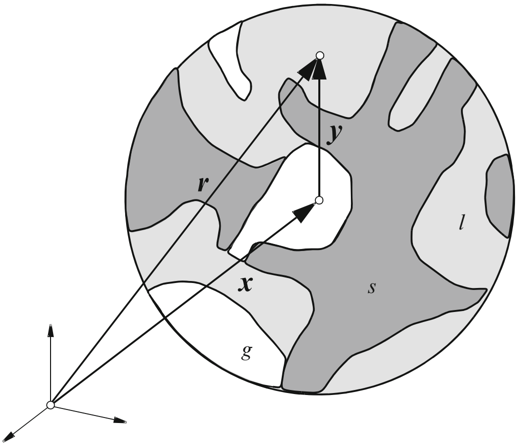


Fig. 3.5 Hypothetical averaging volume *dV* with three phases present

where $dv(\boldsymbol{y})$ is the miscroscopic differential volume and *dV* is the macroscopic REV volume. Clearly, ε_α is constrained by

$$\sum_\alpha \varepsilon_\alpha = 1 \quad \text{and} \quad 0 \le \varepsilon_\alpha \le 1 \tag{3.5}$$

For the macroscopization process we need three different averaging operators:
Volume average operator

$$\langle\,\rangle_\alpha (\boldsymbol{x}, t) = \frac{1}{dV} \int_{dV} (\,)\, \gamma_\alpha(\boldsymbol{x} + \boldsymbol{y}, t) dv(\boldsymbol{y}) \tag{3.6}$$

Intrinsic volume average operator

$$\langle\,\rangle^\alpha (\boldsymbol{x}, t) = \frac{1}{dV_\alpha(\boldsymbol{x}, t)} \int_{dV} (\,)\, \gamma_\alpha(\boldsymbol{x} + \boldsymbol{y}, t) dv(\boldsymbol{y}) \tag{3.7}$$

Intrinsic mass average (Boltzmann) operator

$$\overline{(\,)}^\alpha (\boldsymbol{x}, t) = \frac{1}{\langle\rho\rangle_\alpha dV} \int_{dV} (\,)\, \rho(\boldsymbol{x} + \boldsymbol{y}, t) \gamma_\alpha(\boldsymbol{x} + \boldsymbol{y}, t) dv(\boldsymbol{y}) \tag{3.8}$$

From (3.6) and (3.7), it follows that the volume average and the intrinsic volume average of a scalar quantity ψ are related to each other by

$$\langle\psi\rangle_\alpha = \varepsilon_\alpha \langle\psi\rangle^\alpha \tag{3.9}$$

where ε_α is defined by (3.4). Furthermore, it results from (3.8)

$$\langle\rho\rangle_\alpha \overline{\psi}^\alpha = \langle\rho\psi\rangle_\alpha \tag{3.10}$$

The *deviation* of a microscopic quantity ψ at the point $\boldsymbol{r}$ from its mass average of $\alpha-$phase at the point $\boldsymbol{x}$ is denoted by the *fluctuation* $\tilde{\psi}^{\alpha}$:

$$\tilde{\psi}^{\alpha}(\boldsymbol{x}, \boldsymbol{y}, t) = \psi(\boldsymbol{x} + \boldsymbol{y}, t) - \overline{\psi}^{\alpha}(\boldsymbol{x}, t) \tag{3.11}$$

or

$$\psi = \overline{\psi}^{\alpha} + \tilde{\psi}^{\alpha} \tag{3.12}$$

Since $\overline{\psi}^{\alpha}(\boldsymbol{x}, t)$ is constant in the REV, dV, the following holds:

$$\begin{aligned} \overline{\tilde{\psi}^{\alpha}}^{\alpha} &= 0 \\ \overline{\tilde{\psi}^{\alpha}\, \overline{\phi}^{\alpha}}^{\alpha} &= \overline{\tilde{\psi}^{\alpha}}^{\alpha}\, \overline{\phi}^{\alpha} = 0 \\ \overline{\psi\phi}^{\alpha} &= \overline{\psi}^{\alpha}\, \overline{\phi}^{\alpha} + \overline{\tilde{\psi}^{\alpha}\, \tilde{\phi}^{\alpha}}^{\alpha} \end{aligned} \tag{3.13}$$

where $\phi = \phi(\boldsymbol{r}, t)$ is another scalar quantity.

3.4 Averaging Theorems

Averaging differential expressions within the REV, we have to consider terms providing averages of derivations with respect to space and time. The following theorems relate the average of a gradient and a time derivative to the gradient and time derivative of an average, respectively. For an extensive quantity $\mathcal{F}$, cf. (2.116), it is valid:
Averaging theorem

$$\langle \nabla \cdot \mathcal{F} \rangle_{\alpha} = \nabla \cdot \langle \mathcal{F} \rangle_{\alpha} + \frac{1}{dV} \sum_{\beta \neq \alpha} \int_{dA_{\alpha\beta}} \mathcal{F} \cdot \boldsymbol{n}^{\alpha\beta} da(\boldsymbol{y}) \tag{3.14}$$

Transport theorem

$$\left\langle \frac{\partial \mathcal{F}}{\partial t} \right\rangle_{\alpha} = \frac{\partial}{\partial t} \langle \mathcal{F} \rangle_{\alpha} - \frac{1}{dV} \sum_{\beta \neq \alpha} \int_{dA_{\alpha\beta}} \mathcal{F} \cdot (\boldsymbol{w} \cdot \boldsymbol{n}^{\alpha\beta})\, da(\boldsymbol{y}) \tag{3.15}$$

where $dA_{\alpha\beta}$ is the macroscopic differential interface between $\alpha-$phase and $\beta-$phase within dV, $da(\boldsymbol{y})$ is an elemental portion of this area, $\boldsymbol{n}^{\alpha\beta} = -\boldsymbol{n}^{\beta\alpha}$ is a normal direction vector on this surface pointing from the $\alpha-$phase toward the $\beta-$phase, $\boldsymbol{w}$ is the $\alpha\beta-$interface velocity and ∇ is regarded as the macroscopic gradient operator with respect to the macroscopic coordinates $\boldsymbol{x}$. We note that $\mathcal{F}$ can be either vectorial or scalar quantities.

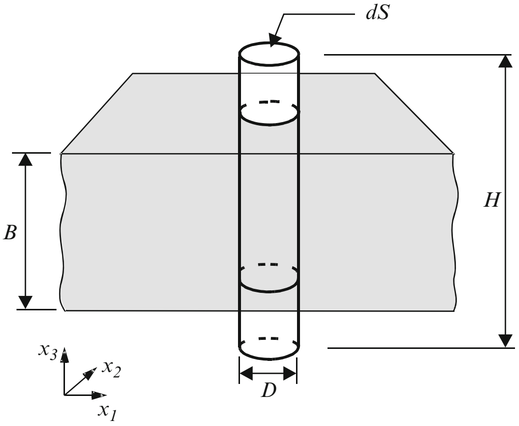


Fig. 3.6 Aquifer averaging volume (AREV) of height H and diameter D penetrating an aquifer of thickness B

3.5 Aquifer Averaging

Flow and transport process modeling in aquifers possesses important special cases, in which the horizontal extent of a regional flow field can be much bigger than the thickness B of an aquifer. For such conditions vertical variations can often be neglected to reduce the full 3D equations to two-dimensional (2D) essentially horizontal relationships. Regarding groundwater hydraulic processes this procedure is associated with the well-known *Dupuit assumption* [33].

There are two distinctly different approaches to develop the macroscopic, 2D relations for aquifers. The standard two-step averaging procedure takes in a first step the above REV averaging technique to derive the general 3D macroscopic equations. In a second step, these equations have to be vertically integrated or vertically averaged. The difficulty with the two-step averaging procedure is that a number of terms whose physical meaning is not readily apparent arise involving derivations from averages which must be taken into account. In contrast, a more general and physically rigorous averaging technique has been proposed by Gray [202], which represents a one-step averaging procedure. It allows a direct transformation from 3D microscopic equations to 2D macroscopic aquifer-related equations. The procedure is termed as *aquifer averaging* and represents an extension to the REV concept.

Aquifer averaging is based on an aquifer REV, termed as AREV, as shown in Fig. 3.6. Within the AREV the following constraint must be satisfied in addition

$$H > B \gg D \tag{3.16}$$

where B is the thickness of the aquifer, which may vary in space and time, and H is the total length of the cylinder, which is constant. In Fig. 3.6, dS denotes the projected circular planar area of the averaging volume, $\pi D^2/4$. In the AREV the position vector $\boldsymbol{r}$ only lies in the horizontal plane. The vertical direction is treated explicitly and referred to as the x_3-direction. Thus, a point in the AREV may be located by specification of its $(\boldsymbol{r}, x_3)$ coordinates.

In modification of the REV procedures, the AREV conception uses the following modified averaging operators, quantities and theorems:
Volume fraction

$$\varepsilon_\alpha(\boldsymbol{x},t) = \frac{dV_\alpha(\boldsymbol{x},t)}{B(\boldsymbol{x},t)\ dS} = \frac{1}{B(\boldsymbol{x},t)\ dS}\int_{dV} \gamma_\alpha(\boldsymbol{x}+\boldsymbol{y},x_3,t)dv(\boldsymbol{y},x_3) \tag{3.17}$$

Volume average operator

$$\langle\ \rangle_\alpha(\boldsymbol{x},t) = \frac{1}{B(\boldsymbol{x},t)\ dS}\int_{dV} (\,)\,\gamma_\alpha(\boldsymbol{x}+\boldsymbol{y},x_3,t)dv \tag{3.18}$$

Intrinsic volume average operator

$$\langle\ \rangle^\alpha(\boldsymbol{x},t) = \frac{1}{B(\boldsymbol{x},t)\ dS_\alpha(\boldsymbol{x},t)}\int_{dV} (\,)\,\gamma_\alpha(\boldsymbol{x}+\boldsymbol{y},x_3,t)dv \tag{3.19}$$

Intrinsic mass average (Boltzmann) operator

$$\overline{(\,)}^\alpha(\boldsymbol{x},t) = \frac{1}{\langle\rho\rangle_\alpha\ B(\boldsymbol{x},t)\ dS}\int_{dV} (\,)\,\rho(\boldsymbol{x}+\boldsymbol{y},x_3,t)\gamma_\alpha(\boldsymbol{x}+\boldsymbol{y},x_3,t)dv \tag{3.20}$$

Averaging theorem

$$\begin{aligned}\langle\nabla\cdot\mathcal{F}\rangle_\alpha = \frac{1}{B}\nabla\cdot[B\ \ \langle\mathcal{F}\rangle_\alpha] + \frac{1}{B\ dS}\sum_{\beta\neq\alpha}\int_{dA_{\alpha\beta}} \mathcal{F}\cdot\boldsymbol{n}^{\alpha\beta}da\\ +\frac{1}{B\ dS}\int_{dS_\alpha^{\mathrm{TB}}} \mathcal{F}\cdot\boldsymbol{n}^{\mathrm{TB}}da\end{aligned} \tag{3.21}$$

Transport theorem

$$\begin{aligned}\left\langle\frac{\partial\mathcal{F}}{\partial t}\right\rangle_\alpha = \frac{1}{B}\frac{\partial}{\partial t}[B\ \ \langle\mathcal{F}\rangle_\alpha] - \frac{1}{B\ dS}\sum_{\beta\neq\alpha}\int_{dA_{\alpha\beta}} \mathcal{F}\cdot(\boldsymbol{w}\cdot\boldsymbol{n}^{\alpha\beta})\ da\\ -\frac{1}{B\ dS}\int_{dS_\alpha^{\mathrm{TB}}} \mathcal{F}\cdot(\boldsymbol{w}\cdot\boldsymbol{n}^{\mathrm{TB}})\ da\end{aligned} \tag{3.22}$$

in which $\boldsymbol{n}^{\mathrm{TB}}$ is the outward-directed unit normal at the top and bottom of the aquifer. We note that the gradient operator ∇ in (3.21) is only 2D, as there is no vertical gradient of B or $\langle\mathcal{F}\rangle_\alpha$.

Table 3.1 Extensive quantities $\mathcal{F}_k$ and intensive quantities f_k and ψ_k related to species k (no summation over k)

Quantity	$\mathcal{F}_k(t)$	$f_k(\boldsymbol{x},t)$	$\psi_k(\boldsymbol{x},t)$
Mass	$\mathcal{M}_k$	ρ_k	1
Momentum	$\mathcal{V}_k$	$\rho_k\, \boldsymbol{v}_k$	$\boldsymbol{v}_k$
Energy	$\mathcal{E}_k + \mathcal{K}_k$	$\rho_k\, (E_k + \frac{1}{2}v_k^2)$	$E_k + \frac{1}{2}v_k^2$
Entropy	$\mathcal{S}_k$	$\rho_k\, S_k$	S_k

3.6 Fundamental Microscopic Balance Laws and Conservation Principles

The core of the mathematical modeling is formed by the four fundamental physical principles of

- $\mathcal{M}_k$, mass balance,
- $\mathcal{V}_k$, momentum balance,
- $\mathcal{E}_k + \mathcal{K}_k$, total energy balance (*first law of thermodynamics*), and
- $\mathcal{S}_k$, entropy balance

associated with species k. Mass, motion, energy and entropy-related quantities can be defined in a 'microscopic' (single phase) volume element (continuum), for which balance laws are postulated. Mass, momentum, internal (thermal) energy, kinetic energy and entropy, respectively, represent *extensive quantities* $\mathcal{F}_k \in (\mathcal{M}_k, \mathcal{V}_k, \mathcal{E}_k, \mathcal{K}_k, \mathcal{S}_k)$ of species k (i.e., those quantities are additive over volumes), cf. (2.116). *Intensive quantities* f_k concern densities of these extensive properties being independent of the balance volume in form of mass densities, momentum densities, energy densities and entropy densities. In this context ρ_k is introduced as a mass density function and ψ_k as an intensive balance quantity. In accordance with (2.116) and Table 2.1, for an arbitrary volume Ω it is

$$\mathcal{F}_k(t) = \int_\Omega f_k(\boldsymbol{x},t)\, d\Omega = \int_\Omega \rho_k \psi_k(\boldsymbol{x},t)\, d\Omega \tag{3.23}$$

where the intensive balance quantities ψ_k are specified in Table 3.1 for the different extensive quantities. In referring to a spatially fixed *Eulerian* coordinate system, the postulate of balance of the extensive quantity $\mathcal{F}_k$ is stated as:

$$\frac{D\mathcal{F}_k}{Dt} \equiv \frac{D}{Dt}\int_\Omega \rho_k \psi_k \; d\Omega - \int_\Omega \rho_k F_k \; d\Omega = \int_\Omega \rho_k G_k \; d\Omega \tag{3.24}$$

where F_k corresponds to an external production (supply) and G_k corresponds to a net rate of production of $\mathcal{F}_k$. The material derivative (2.38) in (3.24) for the Eulerian description is given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + (\boldsymbol{v}_k^{\mathcal{F}} \cdot \nabla) \tag{3.25}$$

Table 3.2 Diffusive fluxes $\boldsymbol{j}_k^{\mathcal{F}}$ related to species k (no summation over k)

Quantity	$\mathcal{F}_k$	ψ_k	$\boldsymbol{j}_k^{\mathcal{F}}$
Mass	$\mathcal{M}_k$	1	$\mathbf{0}$
Momentum	$\mathcal{V}_k$	$\boldsymbol{v}_k$	$\boldsymbol{\sigma}_k$
Internal energy	$\mathcal{E}_k$	E_k	$\boldsymbol{j}_k^{\mathcal{E}}$
Entropy	$\mathcal{S}_k$	S_k	$\boldsymbol{j}_k^{\mathcal{S}}$

where $\boldsymbol{v}_k^{\mathcal{F}}$ represents the velocity vector of the considered particle associated with the quantity $\mathcal{F}_k$. The general balance statement (3.24) can be expressed by using the Reynolds' transport theorem (2.83) as follows:

$$\frac{D}{Dt}\int_{\Omega} \rho_k \psi_k \, d\Omega = \int_{\Omega} \Big(\frac{D(\rho_k \psi_k)}{Dt} + \rho_k \psi_k (\nabla \cdot \boldsymbol{v}_k^{\mathcal{F}}) \Big) \, d\Omega =$$
$$\int_{\Omega} \Big[\frac{\partial(\rho_k \psi_k)}{\partial t} + \nabla \cdot (\rho_k \psi_k \, \boldsymbol{v}_k^{\mathcal{F}}) \Big] \, d\Omega =$$
$$\int_{\Omega} \rho_k F_k \, d\Omega + \int_{\Omega} \rho_k G_k \, d\Omega \qquad (3.26)$$

Thus, (3.26) can be simply written as

$$\frac{\partial(\rho_k \psi_k)}{\partial t} + \nabla \cdot (\rho_k \psi_k \, \boldsymbol{v}_k^{\mathcal{F}}) = \rho_k (F_k + G_k) \qquad (3.27)$$

because the balance expression becomes independent of the volume Ω in a microscopic description.

The particle velocity $\boldsymbol{v}_k^{\mathcal{F}}$ can be further expressed via a *diffusive law* defined as

$$\boldsymbol{j}_k^{\mathcal{F}} = \rho_k \psi_k (\boldsymbol{v}_k^{\mathcal{F}} - \boldsymbol{v}_k) \qquad (3.28)$$

where $\boldsymbol{j}_k^{\mathcal{F}}$ corresponds to a diffusive flux of species k associated with the extensive quantity $\mathcal{F}_k$. It assumes a linear continuous relation to the particle velocity $\boldsymbol{v}_k$ of species k. By using (3.28) the balance expression (3.27) takes the form

$$\frac{\partial(\rho_k \psi_k)}{\partial t} + \nabla \cdot (\rho_k \psi_k \, \boldsymbol{v}_k) + \nabla \cdot \boldsymbol{j}_k^{\mathcal{F}} = \rho_k (F_k + G_k) \qquad (3.29)$$

The diffusive fluxes $\boldsymbol{j}_k^{\mathcal{F}}$ are summarized in Table 3.2 for the different extensive quantities. Since the particle velocity $\boldsymbol{v}_k$ of species k is generally immeasurable, a diffusive flux

$$\boldsymbol{j}_k = \rho_k (\boldsymbol{v}_k - \boldsymbol{v}) \qquad (3.30)$$

is introduced, which directly relates $\boldsymbol{v}_k$ of species k to the mass-weighted (barycentric) velocity $\boldsymbol{v}$ defined as

$$\boldsymbol{v} = \frac{1}{\rho} \sum_k^N \rho_k \, \boldsymbol{v}_k \tag{3.31}$$

with

$$\rho = \sum_k^N \rho_k \tag{3.32}$$

Accordingly, inserting (3.30) into (3.29) we obtain an appropriate balance expression, viz.,

$$\frac{\partial (\rho_k \psi_k)}{\partial t} + \nabla \cdot (\rho_k \psi_k \, \boldsymbol{v}) + \nabla \cdot (\boldsymbol{j}_k^{\mathcal{F}} + \psi_k \boldsymbol{j}_k) = \rho_k (F_k + G_k) \tag{3.33}$$

where the particle velocity $\boldsymbol{v}_k$ is eliminated, however, in account of specifying the diffusive fluxes $\boldsymbol{j}_k^{\mathcal{F}}$ and $\boldsymbol{j}_k$.

The balance equation (3.33) is still rather general because it implies expressions for each species k. Mostly, however, it is sufficient to specify only balance statements for mass-weighted (barycentric) quantities. In doing so, we sum (3.33) over all species k and obtain

$$\frac{\partial (\rho \psi)}{\partial t} + \nabla \cdot (\rho \psi \, \boldsymbol{v}) + \nabla \cdot \boldsymbol{j} = \rho (F + G) \tag{3.34}$$

where

$$\begin{aligned} \psi &= \tfrac{1}{\rho} \textstyle\sum_k^N \rho_k \psi_k \\ F &= \tfrac{1}{\rho} \textstyle\sum_k^N \rho_k F_k \\ G &= \tfrac{1}{\rho} \textstyle\sum_k^N \rho_k G_k \\ \boldsymbol{j} &= \textstyle\sum_k^N (\boldsymbol{j}_k^{\mathcal{F}} + \psi_k \boldsymbol{j}_k) \end{aligned} \tag{3.35}$$

with $\boldsymbol{j}_k^{\mathcal{M}} = \mathbf{0}$ $(\boldsymbol{v}_k^{\mathcal{M}} \equiv \boldsymbol{v}_k)$ according to the definition (3.28) and finding from (3.30) the identity

$$\sum_k^N \boldsymbol{j}_k = \mathbf{0} \tag{3.36}$$

The barycentric variables ψ, $\boldsymbol{j}$, F and G of the general microscopic balance equation (3.34) are listed in Table 3.3 for the different extensive quantities $\mathcal{F}$ that need to be considered, where E is the barycentric internal energy, S is the

Table 3.3 Microscopic quantities appearing in the general microscopic balance equation (3.34)

Quantity	$\mathcal{F}$	ψ	$\boldsymbol{j}$	F	G
Mass					
barycentric	$\mathcal{M}$	1	$\mathbf{0}$	Q	0
species	$\mathcal{M}_k$	ω_k	$\boldsymbol{j}_k$	r_k/ρ	0
Momentum	$\mathcal{V}$	$\boldsymbol{v}$	$\boldsymbol{\sigma}$	$\boldsymbol{g}$	0
Energy	$\mathcal{E}+\mathcal{K}$	$E+\frac{1}{2}v^2$	$\boldsymbol{j}_T+\boldsymbol{\sigma}\cdot\boldsymbol{v}$	$H+\boldsymbol{g}\cdot\boldsymbol{v}$	0
Entropy	$\mathcal{S}$	S	$\boldsymbol{j}_S$	W	Υ

barycentric entropy, $\boldsymbol{\sigma}$ is the barycentric stress tensor, $\boldsymbol{j}_T$ is the barycentric thermal flux, $\boldsymbol{j}_S$ is the barycentric entropy flux, Q is the barycentric supply of mass, $\boldsymbol{g}$ is the barycentric supply of momentum, H is the barycentric supply of thermal energy, W is the barycentric supply of entropy and Υ is the barycentric entropy production. As seen from Table 3.3 the balance expression (3.34) can be considered as a general microscopic balance equation, where even the balance of species mass (3.33) (at $\boldsymbol{j}_k^{\mathcal{M}} = \mathbf{0}$ and $\psi_k = 1$) can be recognized if we formally set $\psi \to \omega_k$, where $\omega_k = \rho_k/\rho$ is the mass fraction of species k (2.123) and in Table 3.3, $\boldsymbol{j}_k$ accounts for the diffusive flux of species k and r_k accounts for the production rate of species k.

Note that the net rate of production G for mass, momentum and internal energy is zero because these quantities are conserved, i.e., the balance statements for mass, momentum and energy represent *conservation equations*. On the other hand, however, entropy is a non-conservative quantity. The axiom of the *second law of thermodynamics* postulates that the entropy production is always non-negative, i.e.,

$$\rho\Upsilon \geq 0 \tag{3.37}$$

3.7 Macroscopization of Balance Equations

3.7.1 General Balance Equation

The transformation of the microscopic balance equation (3.34) to the macroscopic level uses the averaging procedures of (3.6)–(3.13) in combination with the averaging theorems (3.14) and (3.15) and finally leads to the following general formulation of the macroscopic balance equation written for the $\alpha-$phase:

$$\frac{\partial}{\partial t}(\langle\rho\rangle_\alpha\,\overline{\psi}^\alpha)+\nabla\cdot(\langle\rho\rangle_\alpha\,\overline{\psi}^\alpha\,\overline{\boldsymbol{v}}^\alpha)+\nabla\cdot(\varepsilon_\alpha\boldsymbol{j}^\alpha)-\langle\rho\rangle_\alpha\,[\overline{F}^\alpha+e^\alpha(\rho\psi)+J^\alpha] = \langle\rho\rangle_\alpha\,\overline{G}^\alpha \tag{3.38}$$

where

$$\boldsymbol{j}^{\alpha} = \langle \boldsymbol{j} \rangle^{\alpha} + \langle \rho \rangle^{\alpha} \, \overline{\tilde{\boldsymbol{v}}^{\alpha} \, \tilde{\psi}^{\alpha}}^{\alpha} \tag{3.39}$$

represents the macroscopic non-advective (*dispersive*) flux vector for $\overline{\psi}^{\alpha}$ consisting of the first part of macroscopic *diffusion* $\langle \boldsymbol{j} \rangle_{\alpha}$ and the second part of macroscopic *mechanical dispersion* $\langle \rho \rangle_{\alpha} \, \overline{\tilde{\boldsymbol{v}}^{\alpha} \, \tilde{\psi}^{\alpha}}^{\alpha}$. Furthermore, the term $e^{\alpha}(\rho\psi)$ in (3.38) is

$$e^{\alpha}(\rho\psi) = \frac{1}{\langle \rho \rangle_{\alpha}} \frac{1}{dV} \sum_{\beta \neq \alpha} \int_{dA_{\alpha\beta}} \rho\psi(\boldsymbol{w} - \boldsymbol{v}) \cdot \boldsymbol{n}^{\alpha\beta} \, da \tag{3.40}$$

which describes the exchange of $\overline{\psi}^{\alpha}$ with other phases through phase changes caused by relative motion of phase boundaries. The term J^{α} in (3.38) reads

$$J^{\alpha} = \frac{1}{\langle \rho \rangle_{\alpha}} \frac{1}{dV} \sum_{\beta \neq \alpha} \int_{dA_{\alpha\beta}} \boldsymbol{j} \cdot \boldsymbol{n}^{\alpha\beta} \, da \tag{3.41}$$

which describes the diffusion of $\overline{\psi}^{\alpha}$ across the phase interfaces.

A constraint upon (3.38) may be obtained by summing over all phases α to obtain [226]

$$\sum_{\alpha} \langle \rho \rangle_{\alpha} \left[e^{\alpha}(\rho\psi) + J^{\alpha} \right] = 0 \tag{3.42}$$

assuming that no properties are stored at a phase interface.

For the sake of simplicity the general balance equation (3.38) will be rewritten by omitting the averaging symbols in form of angular brackets and overbars indicating macroscopic quantities. In doing so, we replace the mass density $\langle \rho \rangle_{\alpha}$ by its intrinsic average (3.9), i.e., $\langle \rho \rangle_{\alpha} = \rho_{\alpha} = \varepsilon_{\alpha} \, \langle \rho \rangle^{\alpha} = \varepsilon_{\alpha} \rho^{\alpha}$. It is important to note that we shall designate always throughout the book an *intrinsic* quantity by a phase superscript and a bulk quantity by a phase subscript, e.g., $\psi_{\alpha} = \varepsilon_{\alpha} \psi^{\alpha}$, $Q_{\alpha} = \varepsilon_{\alpha} Q^{\alpha}$ and so forth. Using this convention, we can rewrite (3.38) in the following *divergence* form[1]

$$\frac{\partial}{\partial t}(\varepsilon_{\alpha} \rho^{\alpha} \psi^{\alpha}) + \nabla \cdot (\varepsilon_{\alpha} \rho^{\alpha} \psi^{\alpha} \, \boldsymbol{v}^{\alpha}) + \nabla \cdot (\varepsilon_{\alpha} \boldsymbol{j}^{\alpha}) = \varepsilon_{\alpha} \rho^{\alpha} (F^{\alpha} + F^{\alpha}_{\text{ex}} + G^{\alpha}) \tag{3.43}$$

where

$$F^{\alpha}_{\text{ex}} = e^{\alpha}(\rho\psi) + J^{\alpha} \tag{3.44}$$

[1] It denotes a balance statement in its basic conservation formulation.

Table 3.4 Macroscopic quantities appearing in the general macroscopic balance equation (3.43)

Quantity	$\mathcal{F}^\alpha$	ψ^α	$\boldsymbol{j}^\alpha$	F^α	F^α_{ex}	G^α
Mass						
barycentric	$\mathcal{M}^\alpha$	1	$\mathbf{0}$	Q^α	Q^α_{ex}	0
species	$\mathcal{M}^\alpha_k$	ω^α_k	$\boldsymbol{j}^\alpha_k$	r^α_k/ρ^α	R^α_k/ρ^α	0
Momentum	$\mathcal{V}^\alpha$	$\boldsymbol{v}^\alpha$	$\boldsymbol{\sigma}^\alpha$	$\boldsymbol{g}^\alpha$	$\boldsymbol{f}^\alpha_\sigma$	0
Energy	$\mathcal{E}^\alpha + \mathcal{K}^\alpha$	$E^\alpha + \frac{1}{2}v^{\alpha^2}$	$\boldsymbol{j}^\alpha_T + \boldsymbol{\sigma}^\alpha \cdot \boldsymbol{v}^\alpha$	$H^\alpha + \boldsymbol{g}^\alpha \cdot \boldsymbol{v}^\alpha$	$H^\alpha_{\text{ex}} + \boldsymbol{f}^\alpha_\sigma \cdot \boldsymbol{v}^\alpha$	0
Entropy	$\mathcal{S}^\alpha$	S^α	$\boldsymbol{j}^\alpha_S$	W^α	W^α_{ex}	Υ^α

represents a macroscopic exchange term occurring due to phase change and phase interaction, respectively. The definitions for ψ^α, $\boldsymbol{j}^\alpha$, F^α, F^α_{ex} and G^α in the general macroscopic balance equation (3.43) are listed in Table 3.4 for mass, momentum, energy and entropy balance. If we substitute the balance equation of the barycentric mass $\mathcal{M}^\alpha$ with

$$\frac{\partial}{\partial t}(\varepsilon_\alpha \rho^\alpha) + \nabla \cdot (\varepsilon_\alpha \rho^\alpha\, \boldsymbol{v}^\alpha) = \varepsilon_\alpha \rho^\alpha (Q^\alpha + Q^\alpha_{\text{ex}}) \tag{3.45}$$

in (3.43), we find the *convective* form[2] of the balance equation as

$$\varepsilon_\alpha \rho^\alpha \frac{D^\alpha \psi^\alpha}{Dt} + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}^\alpha) = \varepsilon_\alpha \rho^\alpha [F^\alpha + F^\alpha_{\text{ex}} + G^\alpha - \psi^\alpha (Q^\alpha + Q^\alpha_{\text{ex}})] \tag{3.46}$$

where

$$\frac{D^\alpha \psi^\alpha}{Dt} = \frac{\partial \psi^\alpha}{\partial t} + \boldsymbol{v}^\alpha \cdot \nabla \psi^\alpha \tag{3.47}$$

defines the material derivative of ψ^α of the $\alpha-$phase. The divergence form (3.43) and the convective form (3.46) represent equivalent expressions for the same balance quantity ψ^α. In the following the specific formulations of the macroscopic balance laws are described. Macroscopic conservation equations result for mass, momentum and energy with $G^\alpha = 0$.

3.7.2 Conservation of Mass

For conservation of mass, (3.43) becomes

$$\frac{\partial}{\partial t}(\varepsilon_\alpha \rho^\alpha) + \nabla \cdot (\varepsilon_\alpha \rho^\alpha \boldsymbol{v}^\alpha) = \varepsilon_\alpha \rho^\alpha (Q^\alpha + Q^\alpha_{\text{ex}})$$
$$\text{for} \quad \alpha = s, f \in (l, g) \tag{3.48}$$

[2] It denotes a balance statement in which mass conservation is substituted.

where $\boldsymbol{v}^\alpha$ is the (barycentric) velocity of $\alpha-$phase, Q^α represents the phase-internal supply of mass and Q^α_{ex} accounts for phase change of mass (e.g., ice melting). The conservation of mass requires that the total mass created over all phases must be identical to zero, i.e.,

$$\sum_\alpha \varepsilon_\alpha \rho^\alpha Q^\alpha_{\text{ex}} = 0 \tag{3.49}$$

3.7.3 Conservation of Species Mass

For a chemical species k in the $\alpha-$phase, the mass conservation equation results from (3.43) as

$$\frac{\partial}{\partial t}(\varepsilon_\alpha \rho^\alpha \omega^\alpha_k) + \nabla \cdot (\varepsilon_\alpha \rho^\alpha \omega^\alpha_k \, \boldsymbol{v}^\alpha) + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}^\alpha_k) = \varepsilon_\alpha (r^\alpha_k + R^\alpha_k)$$
$$k = 1, \ldots, N^\alpha \quad \alpha = s, f \in (l, g) \quad \text{for each} \quad k \tag{3.50}$$

written in the divergence form and from (3.46) with Table 3.4 as

$$\varepsilon_\alpha \rho^\alpha \frac{D^\alpha \omega^\alpha_k}{Dt} + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}^\alpha_k) = \varepsilon_\alpha [r^\alpha_k + R^\alpha_k - \rho^\alpha \omega^\alpha_k (Q^\alpha + Q^\alpha_{\text{ex}})]$$
$$k = 1, \ldots, N^\alpha \quad \alpha = s, f \in (l, g) \quad \text{for each} \quad k \tag{3.51}$$

written in the convective form, where $\omega^\alpha_k = \rho^\alpha_k / \rho^\alpha$ is the mass fraction of species k, $\boldsymbol{j}^\alpha_k$ is the dispersive flux of species k, r^α_k is the homogeneous reaction rate of species k and R^α_k is the heterogeneous reaction rate of species k. Equations (3.50) and (3.51) are subject to the restrictions to insure a global conservation of mass:

1. The sum of mass fluxes of all k into phase α must be identical to the total mass change in the $\alpha-$phase, i.e.,

$$\sum_k^{N^\alpha} (r^\alpha_k + R^\alpha_k) = \rho^\alpha (Q^\alpha + Q^\alpha_{\text{ex}}) \tag{3.52}$$

2. The sum of dispersive fluxes of all k vanishes in the $\alpha-$phase, i.e.,

$$\sum_k^{N^\alpha} \boldsymbol{j}^\alpha_k = \boldsymbol{0} \tag{3.53}$$

Taking into account (3.52) and (3.53) and noting that $\sum_k^{N^\alpha} \omega^\alpha_k = 1$, the mass balance equation (3.48) for the phase α is obtained by summing (3.50) over all species k.

The balance laws (3.50) and (3.51), respectively, for species k can be alternatively expressed if introducing the mass concentration $C_k^\alpha = \rho_k^\alpha = \omega_k^\alpha \rho^\alpha$, cf. (2.117), according to

$$\frac{\partial}{\partial t}(\varepsilon_\alpha C_k^\alpha) + \nabla \cdot (\varepsilon_\alpha C_k^\alpha \, \boldsymbol{v}^\alpha) + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}_k^\alpha) = \varepsilon_\alpha (r_k^\alpha + R_k^\alpha)$$
$$k = 1, \ldots, N^\alpha \quad \alpha = s, f \in (l, g) \quad \text{for each} \quad k \tag{3.54}$$

written in the divergence form and

$$\varepsilon_\alpha \rho^\alpha \frac{D^\alpha (C_k^\alpha / \rho^\alpha)}{Dt} + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}_k^\alpha) = \varepsilon_\alpha [r_k^\alpha + R_k^\alpha - C_k^\alpha (Q^\alpha + Q_{\text{ex}}^\alpha)]$$
$$k = 1, \ldots, N^\alpha \quad \alpha = s, f \in (l, g) \quad \text{for each} \quad k \tag{3.55}$$

written in the convective form.

3.7.4 Conservation of Momentum

The momentum equation may be obtained from (3.43) and (3.46) with Table 3.4 in its divergence form

$$\frac{\partial}{\partial t}(\varepsilon_\alpha \rho^\alpha \boldsymbol{v}^\alpha) + \nabla \cdot (\varepsilon_\alpha \rho^\alpha (\boldsymbol{v}^\alpha \boldsymbol{v}^\alpha)) + \nabla \cdot (\varepsilon_\alpha \boldsymbol{\sigma}^\alpha) = \varepsilon_\alpha \rho^\alpha (\boldsymbol{g}^\alpha + \boldsymbol{f}_\sigma^\alpha)$$
$$\text{for} \quad \alpha = s, f \in (l, g) \tag{3.56}$$

and in its convective form

$$\varepsilon_\alpha \rho^\alpha \frac{D^\alpha \boldsymbol{v}^\alpha}{Dt} + \nabla \cdot (\varepsilon_\alpha \boldsymbol{\sigma}^\alpha) = \varepsilon_\alpha \rho^\alpha [\boldsymbol{g}^\alpha + \boldsymbol{f}_\sigma^\alpha - \boldsymbol{v}^\alpha (Q^\alpha + Q_{\text{ex}}^\alpha)]$$
$$\text{for} \quad \alpha = s, f \in (l, g) \tag{3.57}$$

where $\boldsymbol{\sigma}^\alpha$ is the stress tensor of the $\alpha-$phase, $\boldsymbol{g}^\alpha$ is the $\alpha-$phase external supply of momentum due to gravity (and electric or magnetic force fields) and $\boldsymbol{f}_\sigma^\alpha$ is the interfacial drag term, which accounts for the exchange of momentum between the $\alpha-$phase and all other phases due to mechanical interaction and exchange of mass. Note that the material derivative for $\boldsymbol{v}^\alpha$ is

$$\frac{D^\alpha \boldsymbol{v}^\alpha}{Dt} = \frac{\partial \boldsymbol{v}^\alpha}{\partial t} + \boldsymbol{v}^\alpha \cdot (\nabla \boldsymbol{v}^\alpha) \tag{3.58}$$

3.7.5 Conservation of Energy: First Law of Thermodynamics

The conservation of the total energy (the first law of thermodynamics) is obtained from (3.43) and (3.46) with Table 3.4 after subtraction of $\boldsymbol{v}^\alpha$ dotted with (3.56) and (3.57), respectively, in the divergence form

$$\frac{\partial}{\partial t}(\varepsilon_\alpha \rho^\alpha E^\alpha) + \nabla \cdot (\varepsilon_\alpha \rho^\alpha \boldsymbol{v}^\alpha E^\alpha) + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}_T^\alpha) + \varepsilon_\alpha \boldsymbol{\sigma}^\alpha : \nabla \boldsymbol{v}^\alpha = \varepsilon_\alpha \rho^\alpha (H^\alpha + H_{\text{ex}}^\alpha) \quad \text{for} \quad \alpha = s, f \in (l, g) \tag{3.59}$$

and in the convective form

$$\varepsilon_\alpha \rho^\alpha \frac{D^\alpha E^\alpha}{Dt} + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}_T^\alpha) + \varepsilon_\alpha \boldsymbol{\sigma}^\alpha : \nabla \boldsymbol{v}^\alpha = \varepsilon_\alpha \rho^\alpha \big[H^\alpha + H_{\text{ex}}^\alpha - (E^\alpha - \tfrac{1}{2} v^{\alpha^2})(Q^\alpha + Q_{\text{ex}}^\alpha) \big] \quad \text{for} \quad \alpha = s, f \in (l, g) \tag{3.60}$$

where $\boldsymbol{j}_T^\alpha$ is the $\alpha-$phase heat flux, H^α is the $\alpha-$phase external supply of energy and H_{ex}^α accounts for the exchange of energy between the $\alpha-$phase and all other phases due to mechanical interaction and exchange of mass. The term $\varepsilon_\alpha \boldsymbol{\sigma}^\alpha : \nabla \boldsymbol{v}^\alpha$ in (3.59) and (3.60) represents a *dissipation* term of energy (for fluids it is termed as viscous dissipation) as an irreversible heat source due to internal forces (friction)

$$\varepsilon_\alpha \boldsymbol{\sigma}^\alpha : \nabla \boldsymbol{v}^\alpha \leq 0 \tag{3.61}$$

which is always negative (at the given definition) and produces thermal energy. For porous and fractured media, however, the energy dissipation is usually very small and is often negligible.

3.7.6 Entropy Balance

The entropy balance law is often neglected in derivations of equations for porous and fractured flow simulation. However, it is an important law when developing and proving constitutive relations for material properties. From (3.43) and (3.46) with Table 3.4 the entropy balance is in the divergence form

$$\frac{\partial}{\partial t}(\varepsilon_\alpha \rho^\alpha S^\alpha) + \nabla \cdot (\varepsilon_\alpha \rho^\alpha S^\alpha) + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}_S^\alpha) - \varepsilon_\alpha \rho^\alpha (W^\alpha + W_{\text{ex}}^\alpha) = \varepsilon_\alpha \rho^\alpha \Upsilon^\alpha \quad \text{for} \quad \alpha = s, f \in (l, g) \tag{3.62}$$

and in the convective form

$$\varepsilon_\alpha \rho^\alpha \frac{D^\alpha S^\alpha}{Dt} + \nabla \cdot (\varepsilon_\alpha \boldsymbol{j}_S^\alpha) - \varepsilon_\alpha \rho^\alpha [W^\alpha + W_{\text{ex}}^\alpha - S^\alpha (Q^\alpha + Q_{\text{ex}}^\alpha)] = \varepsilon_\alpha \rho^\alpha \Upsilon^\alpha$$

$$\text{for} \quad \alpha = s, f \in (l, g) \tag{3.63}$$

where $\boldsymbol{j}_S^\alpha$ is the $\alpha-$phase entropy flux, W^α is the $\alpha-$phase external supply of entropy, W_{ex}^α accounts for the exchange of entropy between the $\alpha-$phase and all other phases due to mechanical interaction and exchange of mass and Υ^α is the $\alpha-$phase net production of entropy. Usually, the entropy flux $\boldsymbol{j}_S^\alpha$ is assumed proportional to the heat flux and the dispersive mass flux of chemical species such that [116]:

$$\boldsymbol{j}_S^\alpha = \frac{1}{T^\alpha} (\boldsymbol{j}_T^\alpha - \sum_k^{N^\alpha} \mu_k^\alpha \boldsymbol{j}_k^\alpha) \tag{3.64}$$

where $(0 < T^\alpha < \infty)$ represents the *absolute temperature* of the $\alpha-$phase and μ_k^α is the *chemical potential* of the kth-species in the $\alpha-$phase. Furthermore, it may be assumed that the entropy supply term W^α is proportional to the heat supply term according to

$$W^\alpha = \frac{H^\alpha}{T^\alpha} \tag{3.65}$$

3.7.7 Second Law of Thermodynamics

The *second law of thermodynamics* dictates the sign of net entropy production. According to this axiom, the rate of net entropy production for the multiphase system must be always positive, i.e.,

$$\rho \Upsilon = \sum_\alpha \varepsilon_\alpha \rho^\alpha \Upsilon^\alpha \geq 0 \tag{3.66}$$

Substitution of (3.64) and (3.65) into (3.63) and introduction of the Helmholtz free energy of the $\alpha-$phase

$$A^\alpha = E^\alpha - T^\alpha S^\alpha \tag{3.67}$$

into (3.60), replacement of the dispersive species flux $\nabla \cdot (\varepsilon_\alpha \boldsymbol{j}_k^\alpha)$ by (3.51), followed by elimination of H^α between (3.59) and (3.63) yields:

$$
\begin{aligned}
\varepsilon_\alpha \rho^\alpha T^\alpha \Upsilon^\alpha = & -\varepsilon_\alpha \rho^\alpha \Big[\frac{D^\alpha A^\alpha}{Dt} + S^\alpha \frac{D^\alpha T^\alpha}{Dt} - \sum_k^{N^\alpha} (\mu_k^\alpha \frac{D^\alpha \omega_k^\alpha}{Dt}) \Big] \\
& -\varepsilon_\alpha \Big[\frac{\boldsymbol{j}_T^\alpha}{T^\alpha} \cdot \nabla T^\alpha - \boldsymbol{\sigma}^\alpha : \nabla \boldsymbol{v}^\alpha - T^\alpha \sum_k^{N^\alpha} \boldsymbol{j}_k^\alpha \cdot \nabla (\frac{\mu_k^\alpha}{T^\alpha}) - \sum_k^{N^\alpha} \mu_k^\alpha (r_k^\alpha + R_k^\alpha) \Big] \\
& -\varepsilon_\alpha \rho^\alpha \Big[T^\alpha W_{\text{ex}}^\alpha - H_{\text{ex}}^\alpha + \big(A^\alpha - \tfrac{1}{2} v^{\alpha^2} - \sum_k^{N^\alpha} \mu_k^\alpha \omega_k^\alpha \big)(Q^\alpha + Q_{\text{ex}}^\alpha) \Big]
\end{aligned}
\tag{3.68}
$$

Now division by T^α and summation over all phases yield the *Clausius-Duhem inequality* of the total entropy production for the multiphase system:

$$
\begin{aligned}
\rho \Upsilon = -\sum_\alpha \varepsilon_\alpha \Big\{ & \frac{1}{T^\alpha} \Big[\rho^\alpha \Big(\frac{D^\alpha A^\alpha}{Dt} + S^\alpha \frac{D^\alpha T^\alpha}{Dt} - \sum_k^{N^\alpha} (\mu_k^\alpha \frac{D^\alpha \omega_k^\alpha}{Dt}) + \\
& \boldsymbol{f}_\sigma^\alpha \cdot \boldsymbol{v}^\alpha + \big(A^\alpha - \tfrac{1}{2} v^{\alpha^2} - \sum_k^{N^\alpha} \mu_k^\alpha \omega_k^\alpha \big)(Q^\alpha + Q_{\text{ex}}^\alpha) \Big) + \frac{\boldsymbol{j}_T^\alpha}{T^\alpha} \cdot \nabla T^\alpha \\
& + \boldsymbol{\sigma}^\alpha : \nabla \boldsymbol{v}^\alpha + \sum_k^{N^\alpha} \mu_k^\alpha (r_k^\alpha + R_k^\alpha) \Big] + \sum_k^{N^\alpha} \boldsymbol{j}_k^\alpha \cdot \nabla (\frac{\mu_k^\alpha}{T^\alpha}) \\
& + \rho^\alpha W_{\text{ex}}^\alpha \Big\} \geq 0
\end{aligned}
\tag{3.69}
$$

expressing the second law of thermodynamics for porous media. Note that in (3.69) the interface condition (3.42) for the energy, $\sum_\alpha \varepsilon_\alpha \rho^\alpha (H_{\text{ex}}^\alpha + \boldsymbol{f}_\sigma^\alpha \cdot \boldsymbol{v}^\alpha) = 0$, with (3.44) and Table 3.4, has been used to replace the energy exchange term H_{ex}^α.

3.7.8 Vertically Averaged Aquifer Balance Equations

The aquifer macroscopization of the microscopic balance equation (3.34) uses the specific averaging procedures and theorems (3.17)–(3.22). The following aquifer-averaged balance equation results in the general form written for the $\alpha-$phase:

$$
\begin{aligned}
& \frac{\partial}{\partial t} (B \langle \rho \rangle_\alpha \overline{\psi}^\alpha) + \nabla \cdot (B \langle \rho \rangle_\alpha \overline{\psi}^\alpha \overline{\boldsymbol{v}}^\alpha) + \nabla \cdot (B \varepsilon_\alpha \boldsymbol{j}^\alpha) - B \langle \rho \rangle_\alpha [\overline{F}^\alpha + \\
& \qquad e^\alpha (\rho \psi) + J^\alpha] + a^\alpha (\rho \psi) = B \langle \rho \rangle_\alpha \overline{G}^\alpha
\end{aligned}
\tag{3.70}
$$

where

$$
\begin{aligned}
e^{\alpha}(\rho\psi) &= \frac{1}{\langle\rho\rangle_{\alpha}}\frac{1}{B\ dS}\sum_{\beta\neq\alpha}\int_{dA_{\alpha\beta}}\rho\psi(\boldsymbol{w}-\boldsymbol{v})\cdot\boldsymbol{n}^{\alpha\beta}\,da \\
J^{\alpha} &= \frac{1}{\langle\rho\rangle_{\alpha}}\frac{1}{B\ dS}\sum_{\beta\neq\alpha}\int_{dA_{\alpha\beta}}\boldsymbol{j}\cdot\boldsymbol{n}^{\alpha\beta}\,da \\
a^{\alpha}(\rho\psi) &= \frac{1}{dS}\int_{dS^{\mathrm{TB}}}\gamma_{\alpha}[\rho\psi(\boldsymbol{v}-\boldsymbol{w})-\boldsymbol{j}]\cdot\boldsymbol{n}^{\mathrm{TB}}\,da
\end{aligned}
\tag{3.71}
$$

describing exchange of $\overline{\psi}^{\alpha}$ due to phase change and interphase transport, respectively. The new term $a^{\alpha}(\rho\psi)$ in (3.70) and (3.71) contains the averages of the microscopic production plus apparent production in the 2D plane due to the addition of ψ at the upper and lower boundaries of the aquifer. The dispersion flux vector $\boldsymbol{j}^{\alpha}$ corresponds to expression (3.39) introduced before. We note that the thickness of the aquifer B can vary in space and time. The gradient operator ∇ in (3.70) is only 2D in the horizontal extent of the aquifer. In summing over all phases a constraint is obtained similar to (3.42) which indicates that properties must be conserved when considering interface transport:

$$
\sum_{\alpha} B\,\langle\rho\rangle_{\alpha}\,[e^{\alpha}(\rho\psi)+J^{\alpha}]=0 \tag{3.72}
$$

Introducing again the simplified notation the aquifer-average balance equation (3.70) can be written in the general *divergence* form

$$
\frac{\partial}{\partial t}(B\varepsilon_{\alpha}\rho^{\alpha}\psi^{\alpha})+\nabla\cdot(B\varepsilon_{\alpha}\rho^{\alpha}\psi^{\alpha}\,\boldsymbol{v}^{\alpha})+\nabla\cdot(B\varepsilon_{\alpha}\boldsymbol{j}^{\alpha})=B\varepsilon_{\alpha}\rho^{\alpha}(F^{\alpha}+F_{\mathrm{ex}}^{\alpha}+G^{\alpha}) \tag{3.73}
$$

and in the *convective* form

$$
B\varepsilon_{\alpha}\rho^{\alpha}\frac{D^{\alpha}\psi^{\alpha}}{Dt}+\nabla\cdot(B\varepsilon_{\alpha}\boldsymbol{j}^{\alpha})=B\varepsilon_{\alpha}\rho^{\alpha}[F^{\alpha}+F_{\mathrm{ex}}^{\alpha}+G^{\alpha}-\psi^{\alpha}(Q^{\alpha}+Q_{\mathrm{ex}}^{\alpha})] \tag{3.74}
$$

in which[3]

[3] The equivalence of the area- and volume-averaged fluxes is shown for the interface term $A^{\alpha}(\rho\psi)$ of (3.71), cf. [229]. The volume-averaged flux describes the REV average in the form:

$$
[\langle\rho\rangle_{\alpha}\overline{\psi}^{\alpha}(\boldsymbol{v}^{\alpha}-\boldsymbol{W})-\boldsymbol{j}^{\alpha}]\cdot\boldsymbol{n}^{\mathrm{TB}}=\frac{1}{dV}\Big(\int_{dV^{\mathrm{TB}}}\gamma^{\alpha}[\rho\psi(\boldsymbol{v}-\boldsymbol{w})-\boldsymbol{j}]dv\Big)\cdot\boldsymbol{n}^{\mathrm{TB}}
$$

Let us assume that the interface has a thickness D, the volume integral may be written

$$
\frac{1}{dV}\int_{-D/2}^{D/2}\Big(\int_{dS^{\mathrm{TB}}}\gamma^{\alpha}[\rho\psi(\boldsymbol{v}-\boldsymbol{w})-\boldsymbol{j}]\cdot\boldsymbol{n}^{\mathrm{TB}}da\Big)dl\approx\frac{D}{dV}\int_{dS^{\mathrm{TB}}}\gamma^{\alpha}[\rho\psi(\boldsymbol{v}-\boldsymbol{w})-\boldsymbol{j}]\cdot\boldsymbol{n}^{\mathrm{TB}}da
$$

$$\begin{aligned} F^{\alpha}_{\text{ex}} &= e^{\alpha}(\rho\psi) + J^{\alpha} - \frac{a^{\alpha}(\rho\psi)}{B\varepsilon_{\alpha}\rho^{\alpha}} \\ a^{\alpha}(\rho\psi) &= [\varepsilon_{\alpha}\rho^{\alpha}\psi^{\alpha}(\boldsymbol{v}^{\alpha} - \boldsymbol{W}) - \boldsymbol{j}^{\alpha}] \cdot \boldsymbol{n}^{\text{TB}} \end{aligned} \tag{3.75}$$

where $\boldsymbol{W}$ is the velocity of the macroscopic interface forming the upper and lower boundary of the aquifer. The interface condition for $a^{\alpha}(\rho\psi)$ appearing in (3.75) will be subsequently used to specify BC's at the top and bottom of the aquifer. The definitions for ψ^{α}, $\boldsymbol{j}^{\alpha}$, F^{α}, F^{α}_{ex} and G^{α} in the aquifer-average balance equations (3.73) and (3.74) are listed in Table 3.4 for mass, momentum, energy and entropy balance. The following expressions result:

Conservation of mass

$$\frac{\partial}{\partial t}(B\varepsilon_{\alpha}\rho^{\alpha}) + \nabla \cdot (B\varepsilon_{\alpha}\rho^{\alpha}\boldsymbol{v}^{\alpha}) = B\varepsilon_{\alpha}\rho^{\alpha}(Q^{\alpha} + Q^{\alpha}_{\text{ex}}) \qquad \text{for} \quad \alpha = s, f \in (l, g) \tag{3.76}$$

Conservation of species mass

$$\frac{\partial}{\partial t}(B\varepsilon_{\alpha}\rho^{\alpha}\omega^{\alpha}_{k}) + \nabla \cdot (B\varepsilon_{\alpha}\rho^{\alpha}\omega^{\alpha}_{k}\,\boldsymbol{v}^{\alpha}) + \nabla \cdot (B\varepsilon_{\alpha}\boldsymbol{j}^{\alpha}_{k}) = B\varepsilon_{\alpha}(r^{\alpha}_{k} + R^{\alpha}_{k}) \qquad k = 1, \ldots, N^{\alpha} \quad \alpha = s, f \in (l, g) \quad \text{for each} \quad k \tag{3.77}$$

or

$$B\varepsilon_{\alpha}\rho^{\alpha}\frac{D^{\alpha}\omega^{\alpha}_{k}}{Dt} + \nabla \cdot (B\varepsilon_{\alpha}\boldsymbol{j}^{\alpha}_{k}) = B\varepsilon_{\alpha}[r^{\alpha}_{k} + R^{\alpha}_{k} - C^{\alpha}_{k}(Q^{\alpha} + Q^{\alpha}_{\text{ex}})] \qquad k = 1, \ldots, N^{\alpha} \quad \alpha = s, f \in (l, g) \quad \text{for each} \quad k \tag{3.78}$$

Conservation of momentum

$$\frac{\partial}{\partial t}(B\varepsilon_{\alpha}\rho^{\alpha}\boldsymbol{v}^{\alpha}) + \nabla \cdot (B\varepsilon_{\alpha}\rho^{\alpha}(\boldsymbol{v}^{\alpha}\boldsymbol{v}^{\alpha})) + \nabla \cdot (B\varepsilon_{\alpha}\boldsymbol{\sigma}^{\alpha}) = B\varepsilon_{\alpha}\rho^{\alpha}(\boldsymbol{g}^{\alpha} + \boldsymbol{f}^{\alpha}_{\sigma}) \quad \text{for} \quad \alpha = s, f \in (l, g) \tag{3.79}$$

or

$$B\varepsilon_{\alpha}\rho^{\alpha}\frac{D^{\alpha}\boldsymbol{v}^{\alpha}}{Dt} + \nabla \cdot (B\varepsilon_{\alpha}\boldsymbol{\sigma}^{\alpha}) = B\varepsilon_{\alpha}\rho^{\alpha}[\boldsymbol{g}^{\alpha} + \boldsymbol{f}^{\alpha}_{\sigma} - \boldsymbol{v}^{\alpha}(Q^{\alpha} + Q^{\alpha}_{\text{ex}})] \qquad \text{for} \quad \alpha = s, f \in (l, g) \tag{3.80}$$

where mean values are used to replace the line integral. With $dS = dV/D$ we find finally

$$[\langle\rho\rangle_{\alpha}\overline{\psi}^{\alpha}(\boldsymbol{v}^{\alpha} - \boldsymbol{W}) - \boldsymbol{j}^{\alpha}] \cdot \boldsymbol{n}^{\text{TB}} = \frac{1}{dS}\int_{dS^{\text{TB}}} \gamma^{\alpha}[\rho\psi(\boldsymbol{v} - \boldsymbol{w}) - \boldsymbol{j}] \cdot \boldsymbol{n}^{\text{TB}} da$$

which corresponds to $a^{\alpha}(\rho\psi)$ in (3.71).

Conservation of energy

$$\begin{aligned}\frac{\partial}{\partial t}(B\varepsilon_\alpha\rho^\alpha E^\alpha) + \nabla\cdot(B\varepsilon_\alpha\rho^\alpha\boldsymbol{v}^\alpha E^\alpha) + \nabla\cdot(B\varepsilon_\alpha\boldsymbol{j}_T^\alpha) + B\varepsilon_\alpha\boldsymbol{\sigma}^\alpha{:}\nabla\boldsymbol{v}^\alpha =&\\ B\varepsilon_\alpha\rho^\alpha(H^\alpha + H_{\text{ex}}^\alpha) \quad \text{for} \quad \alpha = s, f \in (l, g)&\end{aligned} \tag{3.81}$$

or

$$\begin{aligned}B\varepsilon_\alpha\rho^\alpha\frac{D^\alpha E^\alpha}{Dt} + \nabla\cdot(B\varepsilon_\alpha\boldsymbol{j}_T^\alpha) + B\varepsilon_\alpha\boldsymbol{\sigma}^\alpha{:}\nabla\boldsymbol{v}^\alpha = B\varepsilon_\alpha\rho^\alpha\big[H^\alpha + H_{\text{ex}}^\alpha -&\\ (E^\alpha - \tfrac{1}{2}v^{\alpha^2})(Q^\alpha + Q_{\text{ex}}^\alpha)\big] \quad \text{for} \quad \alpha = s, f \in (l, g)&\end{aligned} \tag{3.82}$$

Balance of entropy

$$\begin{aligned}\frac{\partial}{\partial t}(B\varepsilon_\alpha\rho^\alpha S^\alpha) + \nabla\cdot(B\varepsilon_\alpha\rho^\alpha S^\alpha) + \nabla\cdot(B\varepsilon_\alpha\boldsymbol{j}_S^\alpha) - B\varepsilon_\alpha\rho^\alpha(W^\alpha + W_{\text{ex}}^\alpha) =&\\ B\varepsilon_\alpha\rho^\alpha\Upsilon^\alpha \quad \text{for} \quad \alpha = s, f \in (l, g)&\end{aligned} \tag{3.83}$$

or

$$\begin{aligned}B\varepsilon_\alpha\rho^\alpha\frac{D^\alpha S^\alpha}{Dt} + \nabla\cdot(B\varepsilon_\alpha\boldsymbol{j}_S^\alpha) - B\varepsilon_\alpha\rho^\alpha[W^\alpha + W_{\text{ex}}^\alpha - S^\alpha(Q^\alpha + Q_{\text{ex}}^\alpha)] =&\\ B\varepsilon_\alpha\rho^\alpha\Upsilon^\alpha \quad \text{for} \quad \alpha = s, f \in (l, g)&\end{aligned} \tag{3.84}$$

Second law of thermodynamics

$$\begin{aligned}B\rho\Upsilon = -\sum_\alpha B\varepsilon_\alpha\Bigg\{\frac{1}{T^\alpha}\Big[\rho^\alpha\Big(\frac{D^\alpha A^\alpha}{Dt} + S^\alpha\frac{D^\alpha T^\alpha}{Dt} - \sum_k^{N^\alpha}(\mu_k^\alpha\frac{D^\alpha\omega_k^\alpha}{Dt}) +&\\ \boldsymbol{f}_\sigma^\alpha\cdot\boldsymbol{v}^\alpha + \big(A^\alpha - \tfrac{1}{2}v^{\alpha^2} - \sum_k^{N^\alpha}\mu_k^\alpha\omega_k^\alpha\big)(Q^\alpha + Q_{\text{ex}}^\alpha)\Big) + \frac{\boldsymbol{j}_T^\alpha}{T^\alpha}\cdot\nabla T^\alpha&\\ +\boldsymbol{\sigma}^\alpha{:}\nabla\boldsymbol{v}^\alpha + \sum_k^{N^\alpha}\mu_k^\alpha(r_k^\alpha + R_k^\alpha)\Big] + \sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\cdot\nabla(\frac{\mu_k^\alpha}{T^\alpha})&\\ +\rho^\alpha W_{\text{ex}}^\alpha\Bigg\} \geq 0&\\ \text{for} \quad \alpha = s, f \in (l, g)&\end{aligned} \tag{3.85}$$

3.8 Constitutive Theory

3.8.1 Kinematics

As discussed in Sect. 3.2, a porous medium can be viewed as a body which consists of a number of coexistent continua (phases), one solid phase s and two (or more) fluid phases $f = l, g$. Each phase possesses a reference configuration at time $t = 0$, which will be altered by its motion. The motions of the phases are independent. For a solid phase s in particular, a typical particle which occupies a position $\boldsymbol{X}^s$ at time $t = 0$ may be carried to a new position $\boldsymbol{x}$ at time t. Then, the solid phase motion is given by a displacement function $\boldsymbol{u}^s(\boldsymbol{X}^s, t)$ such that (cf. Sect. 2.1.4):

$$\boldsymbol{x} = \boldsymbol{u}^s(\boldsymbol{X}^s, t) \qquad x_i = u_i^s(X_I^s, t) \quad (i, I = 1, 2, 3) \tag{3.86}$$

Note that the lower case latin index i refers to the deformed position (i.e., spatial coordinates) and the upper case latin index I refers to the reference position (i.e., material coordinates). It is assumed that the inverse of (3.86) exists such that:

$$\boldsymbol{X}^s = (\boldsymbol{u}^s)^{-1}(\boldsymbol{x}, t) \qquad X_I^s = (u_I^s)^{-1}(x_i, t) \quad (i, I = 1, 2, 3) \tag{3.87}$$

To have this mapping continuous and bijective at all times, the Jacobian J^s of this motion must be non-zero and strictly positive (cf. Sect. 2.1.4), i.e.,

$$J^s = \det \boldsymbol{J}^s > 0 \qquad \boldsymbol{J}^s = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{X}^s} = \frac{\partial x_i}{\partial X_I^s} \quad (i, I = 1, 2, 3) \tag{3.88}$$

where $\boldsymbol{J}^s$ represents the deformation tensor of the solid phase s.

With the deformation of the solid phase there is a differential change of the volume dV^s occupied by the particle of the porous solid. This can be expressed by the Jacobian of the deformed and the reference (non-deformed) solid volumes:

$$J^s = \frac{dV^s(\boldsymbol{x}, t)}{dV_0^s(\boldsymbol{X}^s, 0)} \tag{3.89}$$

Due to mass conservation the following must be valid

$$\int_{dV^s} (\varepsilon_s \rho^s)\, dV = \int_{dV_0^s} (\varepsilon_s \rho^s)_0\, dV \tag{3.90}$$

and accordingly

$$\int_{dV_0^s} \left[(\varepsilon_s \rho^s)_0 - (\varepsilon_s \rho^s)\, J^s \right] dV = 0 \tag{3.91}$$

and finally

$$J^s = \frac{(\varepsilon_s \rho^s)_0}{(\varepsilon_s \rho^s)} \tag{3.92}$$

Because $(\varepsilon_s \rho^s)_0$ does not depend on time or spatial coordinate $\boldsymbol{x}$, substitution of (3.92) into (3.48) yields

$$\begin{aligned} \frac{D^s J^s}{Dt} &= J^s[\nabla \cdot \boldsymbol{v}^s - (Q^s + Q^s_{\text{ex}})] \\ &= J^s\, \boldsymbol{\delta} : \left[\boldsymbol{d}^s - \boldsymbol{\delta}\frac{(Q^s + Q^s_{\text{ex}})}{3}\right] \end{aligned} \tag{3.93}$$

where

$$\boldsymbol{d}^s = \tfrac{1}{2}\left[\nabla \boldsymbol{v}^s + (\nabla \boldsymbol{v}^s)^T\right] \qquad d^s_{ij} = \tfrac{1}{2}\left(\frac{\partial v^s_i}{\partial x_j} + \frac{\partial v^s_j}{\partial x_i}\right) \tag{3.94}$$

is the symmetric *rate of deformation tensor* of the solid phase s, $D^s/Dt = \partial/\partial t + (\boldsymbol{v}^s \cdot \nabla)$ is the material derivative for the solid phase and $\boldsymbol{\delta}$ is the Kronecker symbol (2.7). The second-order tensor $\boldsymbol{d}^s$ is sometimes called *rate of strain tensor* appropriate for small deformations. The velocity of the solid phase is defined as the material time rate of change of solid phase motion (2.39):

$$\boldsymbol{v}^s = \boldsymbol{v}^s(\boldsymbol{x}, t) = \dot{\boldsymbol{u}}^s = \frac{\partial \boldsymbol{u}^s(\boldsymbol{X}^s, t)}{\partial t}\Big|_{\boldsymbol{X}^s} \tag{3.95}$$

where $|_{\boldsymbol{X}^s}$ indicates that $\boldsymbol{X}^s$ is held constant. The *strain tensor* of the solid phase is commonly defined as

$$\boldsymbol{e}^s = \tfrac{1}{2}\left[\nabla \boldsymbol{u}^s + (\nabla \boldsymbol{u}^s)^T\right] \tag{3.96}$$

where the relation $D^s\boldsymbol{e}^s/Dt = \boldsymbol{d}^s$ holds. Note that the second-order strain tensor (3.96) is symmetric, i.e., $\boldsymbol{e}^s = \boldsymbol{e}^{s^T}$. This symmetry means that there are six rather than nine independent strains, as might be expected in a 3 × 3 matrix. For convenience it is conventional to arrange the strain components in a vector form termed as *strain pseudovector* $\boldsymbol{\epsilon}^s$ of the solid phase by using the so-called *Voigt notation*. This strain pseudovector $\boldsymbol{\epsilon}^s$ is related to the displacement $\boldsymbol{u}^s$ of the solid phase by the following relationship with denoted matrix operations written in the Euclidean space $\Re^3$:

$$\underbrace{\boldsymbol{\epsilon}^s}_{(6\times1)} = \underbrace{\boldsymbol{L}}_{(6\times3)} \cdot \underbrace{\boldsymbol{u}^s}_{(3\times1)} \tag{3.97}$$

introducing the symmetric gradient operator

$$\boldsymbol{L} = \begin{pmatrix} \nabla_1 & 0 & 0 \\ 0 & \nabla_2 & 0 \\ 0 & 0 & \nabla_3 \\ \nabla_2 & \nabla_1 & 0 \\ 0 & \nabla_3 & \nabla_2 \\ \nabla_3 & 0 & \nabla_1 \end{pmatrix} \tag{3.98}$$

with the strain pseudovector components

$$\boldsymbol{\epsilon}^s = \begin{pmatrix} \varepsilon_1^s \\ \varepsilon_2^s \\ \varepsilon_3^s \\ \gamma_{12}^s \\ \gamma_{23}^s \\ \gamma_{31}^s \end{pmatrix} \tag{3.99}$$

where $\nabla_1 = \partial/\partial x_1$, $\nabla_2 = \partial/\partial x_2$, $\nabla_3 = \partial/\partial x_3$ and ε_i^s and γ_{ij}^s $(i, j = 1, 2, 3)$ denote the normal strain components and the shear strain components of the solid phase, respectively. Accordingly, the divergence of the solid velocity $\nabla \cdot \boldsymbol{v}^s (= \boldsymbol{\delta}{:}\boldsymbol{d}^s)$ can be expressed by displacements as follows

$$\nabla \cdot \boldsymbol{v}^s = \boldsymbol{m}^T \cdot \frac{\partial \boldsymbol{\epsilon}^s}{\partial t} = \boldsymbol{m}^T \cdot \left(\boldsymbol{L} \cdot \frac{\partial \boldsymbol{u}^s}{\partial t} \right) \tag{3.100}$$

where $\boldsymbol{m}$ is a specific unit vector defined as

$$\boldsymbol{m}^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \tag{3.101}$$

For the subsequent derivations it will be convenient to use the solid phase velocity $\boldsymbol{v}^s$ as a reference velocity. We define the *relative velocity* of the $\alpha-$phase as

$$\boldsymbol{v}^{\alpha s} = \boldsymbol{v}^\alpha - \boldsymbol{v}^s \tag{3.102}$$

and the material derivative of the $\alpha-$phase can be written according to

$$\frac{D^\alpha}{Dt} = \frac{D^s}{Dt} + (\boldsymbol{v}^{\alpha s} \cdot \nabla) \tag{3.103}$$

3.8.2 Constitutive Equations

The balance laws given by Eqs. (3.48), (3.51), (3.56), (3.59), and (3.62) with the relations (3.64) and (3.67) constitute $N + (3 + D)M$ equations with the $N(3 + D) + M(5 + 3D + D^2)$ unknowns which are enumerated below:

$$\begin{array}{ccccccccccccc} \varepsilon_\alpha, & \rho^\alpha, & \boldsymbol{v}^\alpha, & \omega_k^\alpha, & \boldsymbol{j}_k^\alpha, & (r_k^\alpha + R_k^\alpha), & \boldsymbol{\sigma}^\alpha, & \boldsymbol{f}_\sigma^\alpha, & A^\alpha, & \boldsymbol{j}_T^\alpha, & S^\alpha, & T^\alpha, & \mu_k^\alpha \\ (M) & (M) & (DM) & (N) & (DN) & (N) & (D^2M) & (DM) & (M) & (DM) & (M) & (M) & (N) \end{array} \tag{3.104}$$

Quantities which are not listed in (3.104) are considered as known or directly related to these variables. Therefore, to close the systems of balance equations $N(2+D)+M(2+2D+D^2)$ additional constitutive equations are needed, which must account for the material properties of the system and their interrelation. The development of these constitutive equations will be done in a more general way, where we choose a set of independent variables to express the unknowns. The entropy inequality, the objectivity principle and the material symmetries will be utilized in order to restrict the general relationships. The remaining $N(2+D)+M(2+2D+D^2)$ unknowns, chosen as *dependent variables* are members of the set $\{\Psi_j\}$ given below:

$$\{\Psi_{j=1 \text{ to } N(2+D)+M(2+2D+D^2)}\} = \{\boldsymbol{j}_k^\alpha, (r_k^\alpha + R_k^\alpha), \boldsymbol{\sigma}^\alpha, \boldsymbol{f}_\sigma^\alpha, A^\alpha, \boldsymbol{j}_T^\alpha, S^\alpha, \mu_k^\alpha\} \tag{3.105}$$

These variables are not directly measurable and they have to be determined as functions of directly measurable variables, hereby termed *independent variables*. The choice of independent variables is made in accordance with the following axioms [157, 521]:

1. *Principle of equipresence*. A variable present as an independent variable in one constitutive equation should be so present in all.
2. *Principle of coordinate invariance – objectivity principle*. Constitutive equations must be stated by a rule which holds equally in all inertial coordinate systems at any fixed time.
3. *Principle of admissibility*. The constitutive relations do not violate the balance laws or the second law of thermodynamics.

The principle of objectivity requires that a constitutive equation must be unchanged under an orthonormal transformation of the spatial reference frame. As shown in [132, 228] this requirement implies that velocity $\boldsymbol{v}^\alpha$ and velocity gradient $\nabla \boldsymbol{v}^\alpha$ have to be replaced by the relative velocity $\boldsymbol{v}^{fs} = \boldsymbol{v}^f - \boldsymbol{v}^s$ (3.102) and the rate of deformation tensor of the fluid phase $\boldsymbol{d}^f = \frac{1}{2}[\nabla \boldsymbol{v}^f + (\nabla \boldsymbol{v}^f)^T]$, respectively. Finally, the constitutive equations for the dependent variables (3.105) are postulated in terms of the following set $\{\Xi_j\}$ of *independent variables*:

$$\{\Xi_{j=1 \text{ to } N(1+D)+M(3+2D+D^2)-D-1}\} = \{\varepsilon_f, \rho^\alpha, \boldsymbol{v}^{fs}, \boldsymbol{d}^f, \boldsymbol{\epsilon}^s, \omega_k^\alpha, \nabla\omega_k^\alpha, T^\alpha - T_0, \nabla T^\alpha\} \tag{3.106}$$

Note that T_0 represents a *reference temperature*. In (3.106) ε_s is not chosen as independent variables because the volume fraction must sum to unity (3.5), knowledge of ε_f provides ε_s.

The constitutive equations (3.106) are subject to the principle of admissibility. The Coleman and Noll method [94] is used in Appendix B to restrict the functional form of the constitutive variables. For the sake of simplicity the following assumptions are made:

- The phases are considered *ideal* in so far as constitutive variables which account for intraphase processes ($\boldsymbol{\sigma}^\alpha, A^\alpha, \mu^\alpha, S^\alpha \quad \text{for} \quad \alpha = s, f$) depend only on the properties of that phase.

Accordingly, the restrictions obtained via the Coleman and Noll method in Appendix B yields:

$$\begin{aligned}
A^f &= A^f(\rho^f, \omega_k^f, T^f) \\
S^f &= S^f(\rho^f, \omega_k^f, T^f) \\
A^s &= A^s(\varepsilon_f, \rho^s, \boldsymbol{\epsilon}^s, \omega_k^s, T^s) \\
S^s &= S^s(\varepsilon_f, \rho^s, \boldsymbol{\epsilon}^s, \omega_k^s, T^s) \\
\mu_k^\alpha &= \mu_k^\alpha(\rho^\alpha, \omega_k^\alpha, T^\alpha) \\
\frac{\partial A^\alpha}{\partial \omega_k^\alpha} &= \mu_k^\alpha \qquad\qquad (\alpha, \beta = s, f) \\
\frac{\partial A^\alpha}{\partial T^\alpha} &= -S^\alpha \\
\rho^s \frac{\partial A^s}{\partial \boldsymbol{\epsilon}^s} &= -\boldsymbol{\sigma}^s + \underbrace{\rho^{s^2} \frac{\partial A^s}{\partial \rho^s}}_{p^s}
\end{aligned} \tag{3.107}$$

Note that the dependence of A^s and S^s on ε_f must remain explicit because ε_s is not adopted as an independent constitutive variable, however, related directly via the unity $\varepsilon_s = 1 - \varepsilon_f$, (3.5). Taking into account (3.107) and introducing the *thermodynamic pressure* of the fluid phases f and of the solid phase s, respectively, according to

$$\begin{aligned}
p^f &= p^f(\rho^f, \omega_k^f, T^f) = \rho^{f^2} \frac{\partial A^f}{\partial \rho^f} \\
p^s &= p^s(\varepsilon_f, \rho^s, \boldsymbol{\epsilon}^s, \omega_k^s, T^s) = \rho^{s^2} \frac{\partial A^s}{\partial \rho^s}
\end{aligned} \tag{3.108}$$

the entropy inequality (B.5) takes the form:

$$\begin{aligned}\rho\Upsilon = \sum_f \boldsymbol{v}^{fs}\cdot\Big\{\frac{1}{T^f}p^f\nabla\varepsilon_f - \frac{\varepsilon_f\rho^f}{T^f}\boldsymbol{f}_\sigma^f\Big\} + \sum_f \boldsymbol{d}^f{:}\Big\{\frac{\varepsilon_f}{T^f}(p^f\boldsymbol{\delta}-\boldsymbol{\sigma}^f)\Big\}\\ -\sum_\alpha \nabla T^\alpha\cdot\Big\{\frac{\varepsilon_\alpha}{T^{\alpha 2}}(\boldsymbol{j}_T^\alpha - \sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\mu_k^\alpha) + \frac{\varepsilon_\alpha}{T^\alpha}\sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\frac{\partial\mu_k^\alpha}{\partial T^\alpha}\Big\}\\ -\sum_\alpha\sum_k \nabla\omega_k^\alpha\cdot\Big(\frac{\varepsilon_\alpha}{T^\alpha}\boldsymbol{j}_k^\alpha\frac{\partial\mu_k^\alpha}{\partial\omega_k^\alpha}\Big)\\ -\sum_\alpha\Big\{\varepsilon_\alpha\sum_k^{N^\alpha}\mu_k^\alpha(r_k^\alpha + R_k^\alpha) + \varepsilon_\alpha\rho^\alpha W_{\text{ex}}^\alpha\Big\}\\ -\sum_\alpha\Big\{\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\Big(A^\alpha - \tfrac{1}{2}v^{\alpha s^2} - \sum_k^{N^\alpha}\mu_k^\alpha\omega_k^\alpha\Big) + \frac{\varepsilon_\alpha}{T^\alpha}p^\alpha\Big\}(Q^\alpha + Q_{\text{ex}}^\alpha)\\ \geq 0\\ \text{for}\quad(\alpha = s, f)\end{aligned}\tag{3.109}$$

3.8.3 Equilibrium Restrictions

Thermodynamic equilibrium is the state where the following independent variables of (3.106) controlling directly the entropy production (3.109)

$$\xi_j \subset \Xi_j = \{\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^\alpha, \nabla T^\alpha\}\tag{3.110}$$

are all zero and the constitutive functions satisfy:

$$\begin{aligned}\textstyle\sum_\alpha\big(\varepsilon_\alpha\sum_k^{N^\alpha}\mu_k^\alpha(r_k^\alpha + R_k^\alpha)\big)\Big|_e &= 0\\ \textstyle\sum_\alpha\big(\varepsilon_\alpha\rho^\alpha W_{\text{ex}}^\alpha\big)\Big|_e &= 0\\ \textstyle\sum_\alpha\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\sum_k^{N^\alpha}\big(\mu_k^\alpha\omega_k^\alpha\big)\Big|_e &= 0\\ \textstyle\sum_\alpha\big(\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}(Q^\alpha + Q_{\text{ex}}^\alpha)G^\alpha\big)\Big|_e &= 0\end{aligned}\tag{3.111}$$

where $\big|_e$ denotes evaluation at the equilibrium and $G^\alpha = A^\alpha + p^\alpha/\rho^\alpha$ is the Gibbs free energy of the $\alpha-$phase.

At the thermodynamic equilibrium the entropy production $\rho\Upsilon$ goes to zero, i.e., it attains its minimum value. The necessary and sufficient conditions to ensure that $\rho\Upsilon$ is a minimum at equilibrium are:

$$\frac{\partial \rho \Upsilon}{\partial \xi_j}\bigg|_e = 0 \quad \text{and} \quad \left\| \frac{\partial \rho \Upsilon}{\partial \xi_i \partial \xi_j} \right\|_e \geq 0 \tag{3.112}$$

Application of restriction (3.112) to (3.109) yields:

$$\begin{aligned} -\varepsilon_f \rho^f \boldsymbol{f}_\sigma^f \big|_e + p^f \nabla \varepsilon_f &= \mathbf{0} \\ -\boldsymbol{\sigma}^f \big|_e + p^f \boldsymbol{\delta} &= \mathbf{0} \\ -\boldsymbol{j}_k^\alpha|_e &= \mathbf{0} \\ -\boldsymbol{j}_T^\alpha|_e + \textstyle\sum_k^{N^\alpha} \boldsymbol{j}_k^\alpha \mu_k^\alpha - T^\alpha \sum_k^{N^\alpha} \boldsymbol{j}_k^\alpha \frac{\partial \mu_k^\alpha}{\partial T^\alpha} &= \mathbf{0} \end{aligned} \tag{3.113}$$

It shows that $\boldsymbol{f}_\sigma^f$ and $\boldsymbol{\sigma}^f$ are composed of an equilibrium part and a non-equilibrium (deviatoric) part, the latter being zero at equilibrium. Accordingly, it is useful to split $\boldsymbol{\sigma}^f$ and $\boldsymbol{f}_\sigma^f$ in such a form

$$\begin{aligned} \boldsymbol{\sigma}^f &= p^f \boldsymbol{\delta} + \boldsymbol{\tau}^f \\ \varepsilon_f \rho^f \boldsymbol{f}_\sigma^f &= p^f \nabla \varepsilon_f + \boldsymbol{f}_\tau^f \end{aligned} \tag{3.114}$$

where $\boldsymbol{\tau}^f$ and $\boldsymbol{f}_\tau^f$ represent the deviatoric fluid stress tensor and the deviatoric fluid momentum exchange vector, respectively. With (3.113) and (3.114) these flux and stress variables dependent on (3.106) ensure at equilibrium:

$$\begin{aligned} \boldsymbol{f}_\tau^f(\varepsilon_f, \rho^\alpha, 0, 0, \boldsymbol{\epsilon}^s, \omega_k^\alpha, 0, T^\alpha - T_0, 0) &= \mathbf{0} \\ \boldsymbol{\tau}^f(\varepsilon_f, \rho^\alpha, 0, 0, \boldsymbol{\epsilon}^s, \omega_k^\alpha, 0, T^\alpha - T_0, 0) &= \mathbf{0} \\ \boldsymbol{j}_k^\alpha(\varepsilon_f, \rho^\alpha, 0, 0, \boldsymbol{\epsilon}^s, \omega_k^\alpha, 0, T^\alpha - T_0, 0) &= \mathbf{0} \\ \boldsymbol{j}_T^\alpha(\varepsilon_f, \rho^\alpha, 0, 0, \boldsymbol{\epsilon}^s, \omega_k^\alpha, 0, T^\alpha - T_0, 0) &= \mathbf{0} \end{aligned} \tag{3.115}$$

As a consequence, if $\boldsymbol{f}_\tau^f$, $\boldsymbol{\tau}^f$, $\boldsymbol{j}_k^\alpha$ and $\boldsymbol{j}_T^\alpha$ will be developed subsequently for the chosen independent variables (3.106) in form of *phenomenological equations*,[4] the equilibrium condition (3.115) requires that dependency can only be allowed for the driving thermodynamic 'forces' $\boldsymbol{v}^{fs}$, $\boldsymbol{d}^f$, $\nabla \omega_k^\alpha$ and ∇T^α, i.e.,

$$\begin{aligned} \boldsymbol{f}_\tau^f &= \boldsymbol{f}_\tau^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla \omega_k^\alpha, \nabla T^\alpha) \\ \boldsymbol{\tau}^f &= \boldsymbol{\tau}^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla \omega_k^\alpha, \nabla T^\alpha) \\ \boldsymbol{j}_k^\alpha &= \boldsymbol{j}_k^\alpha(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla \omega_k^\alpha, \nabla T^\alpha) \\ \boldsymbol{j}_T^\alpha &= \boldsymbol{j}_T^\alpha(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla \omega_k^\alpha, \nabla T^\alpha) \end{aligned} \tag{3.116}$$

[4] They represent constitutive relations originally found from the observation that fluxes of extensive quantities (e.g., mass, heat, momentum) are produced by the nonuniform distribution of their state variables (e.g., concentration gradient, temperature gradient, velocity difference). Frequently, a simple proportionality between fluxes and gradients of state variables is postulated using a parameter taken to be a property of the material (e.g., diffusivity, conductivity, friction).

3.8.4 Basic Balance Equations and Entropy Inequality

The combination of the above constitutive equations with the general balance equations (3.48), (3.51), (3.57), and (3.60) as well as with the entropy inequality (3.109) yields the following relations by using the substitution of bulk source/sink terms $Q_\alpha = \varepsilon_\alpha(Q^\alpha + Q^\alpha_{\text{ex}}), H_\alpha = \varepsilon_\alpha(H^\alpha + H^\alpha_{\text{ex}}), \alpha = f, s$:

Mass conservation of fluid phases

$$\frac{D^f(\varepsilon_f \rho^f)}{Dt} + \varepsilon_f \rho^f (\boldsymbol{\delta}:\boldsymbol{d}^f) - \rho^f Q_f = 0 \qquad \text{for} \quad f = l, g \tag{3.117}$$

Mass conservation of solid phase

$$\frac{D^s(\varepsilon_s \rho^s)}{Dt} + \varepsilon_s \rho^s (\boldsymbol{m}^T \cdot \frac{\partial \boldsymbol{\epsilon}^s}{\partial t}) - \rho^s Q_s = 0 \tag{3.118}$$

Mass conservation of species k of fluid phases

$$\varepsilon_f \rho^f \frac{D^f \omega_k^f}{Dt} + \nabla \cdot (\varepsilon_f \boldsymbol{j}_k^f) - \varepsilon_f (r_k^f + R_k^f) + \rho^f \omega_k^f Q_f = 0 \qquad \text{for} \quad f = l, g \tag{3.119}$$

Mass conservation of species k of solid phase

$$\varepsilon_s \rho^s \frac{D^s \omega_k^s}{Dt} - \varepsilon_s (r_k^s + R_k^s) + \rho^s \omega_k^s Q_s = 0 \tag{3.120}$$

Momentum conservation of fluid phases

$$\varepsilon_f \rho^f \frac{D^f \boldsymbol{v}^f}{Dt} + \varepsilon_f \nabla p^f + \nabla \cdot (\varepsilon_f \boldsymbol{\tau}^f) - \varepsilon_f \rho^f \boldsymbol{g} - \boldsymbol{f}_\tau^f + \rho^f \boldsymbol{v}^f Q_f = \boldsymbol{0} \qquad \text{for} \quad f = l, g \tag{3.121}$$

Momentum conservation of solid phase

$$\varepsilon_s \rho^s \frac{\partial^2 \boldsymbol{u}^s}{\partial t^2} + \nabla \cdot (\varepsilon_s \boldsymbol{\sigma}^s) - \varepsilon_s \rho^s \boldsymbol{g} + \rho^s \boldsymbol{v}^s Q_s = \boldsymbol{0} \tag{3.122}$$

Energy conservation of fluid phases

$$\varepsilon_f \rho^f \frac{D^f E^f}{Dt} + \nabla \cdot (\varepsilon_f \boldsymbol{j}_T^f) + \varepsilon_f (p^f \boldsymbol{\delta} + \boldsymbol{\tau}^f) : \boldsymbol{d}^f - \rho^f H_f + \rho^f (E^f - \tfrac{1}{2} v^{f^2}) Q_f = 0 \quad \text{for} \quad f = l, g \tag{3.123}$$

Energy conservation of solid phase

$$\varepsilon_s \rho^s \frac{D^s E^s}{Dt} + \nabla \cdot (\varepsilon_s \boldsymbol{j}_T^s) + \varepsilon_s \boldsymbol{\sigma}^s : \boldsymbol{d}^s - \rho^s H_s + \rho^s (E^s - \tfrac{1}{2} v^{s^2}) Q_s = 0 \tag{3.124}$$

Entropy of the fluid-solid phase system

$$\begin{aligned} \rho \Upsilon = {} & \sum_f \boldsymbol{v}^{fs} \cdot (-\frac{1}{T^f} \boldsymbol{f}_\tau^f) + \sum_f \boldsymbol{d}^f : (-\frac{\varepsilon_f}{T^f} \boldsymbol{\tau}^f) \\ & - \sum_\alpha \nabla T^\alpha \cdot \Big\{ \frac{\varepsilon_\alpha}{T^{\alpha 2}} (\boldsymbol{j}_T^\alpha - \sum_k^{N^\alpha} \boldsymbol{j}_k^\alpha \mu_k^\alpha) + \frac{\varepsilon_\alpha}{T^\alpha} \sum_k^{N^\alpha} \boldsymbol{j}_k^\alpha \frac{\partial \mu_k^\alpha}{\partial T^\alpha} \Big\} \\ & - \sum_f \sum_k \nabla \omega_k^f \cdot \Big(\frac{\varepsilon_f}{T^f} \boldsymbol{j}_k^f \frac{\partial \mu_k^f}{\partial \omega_k^f} \Big) - \sum_\alpha \Big\{ \varepsilon_\alpha \sum_k^{N^\alpha} \mu_k^\alpha (r_k^\alpha + R_k^\alpha) + \varepsilon_\alpha \rho^\alpha W_{\text{ex}}^\alpha \Big\} \\ & - \sum_\alpha \Big\{ \frac{\rho^\alpha}{T^\alpha} \Big(A^\alpha - \tfrac{1}{2} v^{\alpha s^2} - \sum_k^{N^\alpha} \mu_k^\alpha \omega_k^\alpha \Big) + \frac{1}{T^\alpha} p^\alpha \Big\} Q_\alpha \geq 0 \\ & \text{for} \quad (\alpha = s, f) \end{aligned} \tag{3.125}$$

In the above equations, the following useful assumptions are made:

- The stress tensors $\boldsymbol{\tau}^f$ and $\boldsymbol{\sigma}^s$ are symmetric.
- The only external supply of momentum is provided by the gravity, i.e., $\boldsymbol{g} = \boldsymbol{g}^\alpha$.
- Diffusive (dispersive) flux of chemical species in the solid phase does not exist, i.e., $\boldsymbol{j}_k^s = \boldsymbol{0}$.

Furthermore, the entropy balance (3.63) is not included anymore because we need not to know explicitly the entropy variable S^α in the subsequent analysis. Accordingly, the conservation laws (3.117)–(3.124) provide now $N + M(2 + D)$ equations, which are available for solving:

$$
\begin{array}{llll}
\left.\begin{array}{lll}
\rho^f & (\text{or} \quad p^f) & \text{from (3.117)} \\
\rho^s & (\text{or} \quad p^s) & \text{from (3.118)}
\end{array}\right\} & M & \text{equations} \\
\left.\begin{array}{lll}
\omega_k^f & & \text{from (3.119)} \\
\omega_k^s & & \text{from (3.120)}
\end{array}\right\} & N & \text{equations} \\
\left.\begin{array}{lll}
\boldsymbol{v}^f & & \text{from (3.121)} \\
\boldsymbol{u}^s & & \text{from (3.122)}
\end{array}\right\} & DM & \text{equations} \\
\left.\begin{array}{lll}
E^f & (\text{or} \quad T^f) & \text{from (3.123)} \\
E^s & (\text{or} \quad T^s) & \text{from (3.124)}
\end{array}\right\} & M & \text{equations}
\end{array}
\tag{3.126}
$$

To close this system of equations the following list of dependent variables remains

$$
\{\Psi_{j=1 \text{ to } N(1+D)+D(2M-1)+D^2M}\} = \{\boldsymbol{\tau}^f, \boldsymbol{f}_\tau^f, \boldsymbol{\sigma}^s, \boldsymbol{j}_k^f, \boldsymbol{j}_T^\alpha, (r_k^\alpha + R_k^\alpha)\} \tag{3.127}
$$

which must be determined by appropriate constitutive functions depending suitably on the independent variables (3.106). Furthermore, *equations of state* (EOS) have to be supplemented in order to determine the needed explicit information about the fluid density ρ^f and the internal energy E^α $(\alpha = s, f)$

$$
\begin{aligned}
\rho^f &= \rho^f(p^f, \omega_k^f, T^f) \\
E^f &= E^f(\rho^f, \omega_k^f, T^f) \\
E^s &= E^s(\varepsilon_f, \rho^s, \boldsymbol{\epsilon}^s, \omega_k^s, T^s)
\end{aligned}
\tag{3.128}
$$

taking into account (3.67) and the restrictions (3.107).

3.8.5 Development of Phenomenological Equations and Constitutive Relations

The remaining dependent variables as listed in (3.127) have to be expressed by phenomenological and constitutive functions in terms of the independent variables (3.106):

$$
\begin{aligned}
\boldsymbol{f}_\tau^f &= \boldsymbol{f}_\tau^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^f, \nabla T^\alpha) \\
\boldsymbol{\tau}^f &= \boldsymbol{\tau}^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^f, \nabla T^\alpha) \\
\boldsymbol{\sigma}^s &= \boldsymbol{\sigma}^s(\varepsilon_f, \rho^s, \boldsymbol{\epsilon}^s, \omega_k^s, T^s) \\
\boldsymbol{j}_k^f &= \boldsymbol{j}_k^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^f, \nabla T^\alpha) \\
\boldsymbol{j}_T^\alpha &= \boldsymbol{j}_T^\alpha(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^\alpha, \nabla T^\alpha) \\
(r_k^\alpha + R_k^\alpha) &= (r_k^\alpha + R_k^\alpha)(\rho^f, \omega_k^\alpha, T^\alpha)
\end{aligned}
\tag{3.129}
$$

where the obtained restrictions (3.107), (3.111), and (3.116) have been taken into account. Polynomial expansions are usually employed up to the desired degree of approximation. A truncated Taylor series around the state $\Xi = 0$ leads to expressions as exemplified for $\boldsymbol{j}_T^\alpha$:

$$\boldsymbol{j}_T^\alpha = -\sum_\gamma \frac{\partial \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{\gamma s}}\Big|_0 \cdot \boldsymbol{v}^{\gamma s} - \sum_\gamma \frac{\partial \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{d}^\gamma}\Big|_0 \cdot \boldsymbol{d}^\gamma - \sum_\beta \sum_k^{N^\alpha} \frac{\partial \boldsymbol{j}_T^\alpha}{\partial \nabla \omega_k^\beta}\Big|_0 \cdot \nabla \omega_k^\beta - \sum_\beta \frac{\partial \boldsymbol{j}_T^\alpha}{\partial \nabla T^\beta}\Big|_0 \cdot \nabla T^\beta$$
$$- \text{HOT} \quad \text{for} \quad (\gamma = f)\ (\beta = f, s) \tag{3.130}$$

where $\big|_0$ denotes evaluation at $\Xi = 0$ and HOT represents higher-order terms:

$$\begin{aligned} \text{HOT} = {} & \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}) + \mathcal{O}(\boldsymbol{d}^\gamma \boldsymbol{d}^\gamma) + \mathcal{O}(\nabla \omega_k^\beta \nabla \omega_k^\beta) + \mathcal{O}(\nabla T^\beta \nabla T^\beta) \\ & + \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{d}^\gamma) + \mathcal{O}(\boldsymbol{v}^{\gamma s}\nabla \omega_k^\beta) + \mathcal{O}(\boldsymbol{v}^{\gamma s}\nabla T^\beta) + \ldots + \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}) \\ & + \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}\boldsymbol{d}^\gamma) + \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}\nabla \omega_k^\beta) + \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}\nabla T^\beta) + \ldots \end{aligned} \tag{3.131}$$

in which for instance

$$\begin{aligned} \mathcal{O}(\boldsymbol{v}^{\gamma s}\nabla T^\beta) &= \sum_\gamma \sum_\beta \tfrac{1}{2} \frac{\partial^2 \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{\gamma s} \partial \nabla T^\beta}\Big|_0 \cdot (\boldsymbol{v}^{\gamma s}\nabla T^\beta) \\ \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}\nabla T^\beta) &= \sum_\gamma \sum_\beta \tfrac{1}{6} \frac{\partial^3 \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{\gamma s} \partial \boldsymbol{v}^{\gamma s} \partial \nabla T^\beta}\Big|_0 \cdot (\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}\nabla T^\beta) \end{aligned} \tag{3.132}$$

or written in index notation $(i, j, m, n = 1, \ldots, D)$

$$\begin{aligned} \mathcal{O}(\boldsymbol{v}^{\gamma s}\nabla T^\beta)_i &= \sum_\gamma \sum_\beta \underbrace{\tfrac{1}{2} \frac{\partial^2 j_{Ti}^\alpha}{\partial v_j^{\gamma s} \partial (\partial T^\beta / \partial x_m)}\Big|_0}_{A_{ijm}^{\alpha\gamma\beta}} \left(v_j^{\gamma s} \frac{\partial T^\beta}{\partial x_m} \right) \\ \mathcal{O}(\boldsymbol{v}^{\gamma s}\boldsymbol{v}^{\gamma s}\nabla T^\beta)_i &= \sum_\gamma \sum_\beta \underbrace{\tfrac{1}{6} \frac{\partial^3 j_{Ti}^\alpha}{\partial v_j^{\gamma s} \partial v_m^{\gamma s} \partial (\partial T^\beta / \partial x_n)}\Big|_0}_{B_{ijmn}^{\alpha\gamma\beta}} \left(v_j^{\gamma s} v_m^{\gamma s} \frac{\partial T^\beta}{\partial x_n} \right) \end{aligned} \tag{3.133}$$

In the above Taylor series the derivative terms $(.)\big|_0$ represent tensorial quantities, which account for material properties and have to be known (or to be determined) as *material coefficients*. Note that in (3.130) a negative sign is used for the development. This is required by the entropy inequality (3.125), where $\boldsymbol{j}_T^\alpha$ has to be negative while the material coefficients remain positive. The higher order terms of the material coefficients lead to tensorial coefficients of higher order so as indicated in (3.133), where 3rd-order and 4th-order tensors $A_{ijm}^{\alpha\gamma\beta}$, $B_{ijmn}^{\alpha\gamma\beta}$ appear. If we assume

- Higher order tensors for the $\alpha-$phase are isotropic and symmetric.
- Material coefficients related to the $\alpha-$phase depend only on the properties of that phase, i.e., for instance $\sum_\gamma \sum_\beta A_{ijm}^{\alpha\gamma\beta}(v_j^{\gamma s} \frac{\partial T^\beta}{\partial x_m}) = A_{ijm}^\alpha (v_j^{\alpha s} \frac{\partial T^\alpha}{\partial x_m})$.

then any 3rd-order tensor $A_{ijm}^{\alpha\gamma\beta}$ and any 4th-order tensor $B_{ijmn}^{\alpha\gamma\beta}$ simplify, cf. [12,132]

$$
\begin{aligned}
A^{\alpha\gamma\beta}_{ijm} &\rightarrow A^{\alpha}_{ijm} &&= 0 \quad \text{(no isotropic odd-order tensors exist)} \\
B^{\alpha\gamma\beta}_{ijmn} &\rightarrow B^{\alpha}_{ijmn} &&= b^{\alpha}_0 \delta_{ij}\delta_{mn} + b^{\alpha}_1(\delta_{im}\delta_{jn} + \delta_{in}\delta_{jm})
\end{aligned}
\tag{3.134}
$$

where b^{α}_0 and b^{α}_1 correspond to material coefficients of the $\alpha-$phases. The following derivations will take into account these assumptions.

3.8.5.1 Deviatoric Fluid Stress Tensor $\boldsymbol{\tau}^f$

The Taylor series expansion of $\boldsymbol{\tau}^f = \boldsymbol{\tau}^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega^f_k, \nabla T^{\alpha})$ for the 1st-order terms yields

$$
\boldsymbol{\tau}^f = -\frac{\partial \boldsymbol{\tau}^f}{\partial \boldsymbol{v}^{fs}}\bigg|_0 \cdot \boldsymbol{v}^{fs} - \frac{\partial \boldsymbol{\tau}^f}{\partial \boldsymbol{d}^f}\bigg|_0 \cdot \boldsymbol{d}^f - \sum_{k}^{N^f} \frac{\partial \boldsymbol{\tau}^f}{\partial \nabla\omega^f_k}\bigg|_0 \cdot \nabla\omega^f_k - \frac{\partial\partial \boldsymbol{\tau}^f}{\partial \nabla T^f}\bigg|_0 \cdot \nabla T^f
\tag{3.135}
$$

where HOT are neglected and a negative sign is used due to the entropy restriction $\boldsymbol{\tau}^f \leq 0$ in (3.125). Considering isotropic conditions we find for the 3rd-order tensors:

$$
\frac{\partial \boldsymbol{\tau}^f}{\partial \boldsymbol{v}^{fs}}\bigg|_0 = \frac{\partial \boldsymbol{\tau}^f}{\partial \nabla\omega^f_k}\bigg|_0 = \frac{\partial\partial \boldsymbol{\tau}^f}{\partial \nabla T^f}\bigg|_0 = \mathbf{0}
\tag{3.136}
$$

and for the 4th-order symmetric tensor:

$$
\frac{\partial \boldsymbol{\tau}^f}{\partial \boldsymbol{d}^f}\bigg|_0 = \bar{\tau}^f_{ijmn} = \lambda^f \delta_{ij}\delta_{mn} + \mu^f(\delta_{im}\delta_{jn} + \delta_{in}\delta_{jm})
\tag{3.137}
$$

where λ^f is denoted as dilatational (or bulk) viscosity and μ^f is denoted as dynamic (or shear) viscosity of the $f-$phase. With (3.136) and (3.137) we obtain from (3.135)

$$
\begin{aligned}
\boldsymbol{\tau}^f &= -\big(\lambda^f(\boldsymbol{\delta}:\boldsymbol{d}^f)\boldsymbol{\delta} + 2\mu^f \boldsymbol{d}^f\big) \\
\tau^f_{ij} &= -\big(\lambda^f d^f_{mm}\delta_{ij} + 2\mu^f d^f_{ij}\big)
\end{aligned}
\tag{3.138}
$$

In fluid mechanics the *mechanical pressure* p^f_{mech} is defined as the average of the normal stress

$$
p^f_{\text{mech}} = \tfrac{1}{3}\boldsymbol{\delta}:\boldsymbol{\sigma}^f = \tfrac{1}{3}\sigma^f_{ii}
\tag{3.139}
$$

The difference between the thermodynamic pressure p^f defined by (3.108) and the mechanical pressure p^f_{mech} defined by (3.139) is obtained from (3.114) by contracting on the index i and dividing by 3. It results

$$
p^f - p^f_{\text{mech}} = (\lambda^f + \tfrac{2}{3}\mu^f)\boldsymbol{\delta}:\boldsymbol{d}^f
\tag{3.140}
$$

The assumption that the two pressures are equal is known as *Stokes' assumption*, and it means that

$$\lambda^f = -\tfrac{2}{3}\mu^f \tag{3.141}$$

Then the deviatoric stress tensor $\boldsymbol{\tau}^f$ reaches its final form

$$\boldsymbol{\tau}^f = \tfrac{2}{3}\mu^f(\boldsymbol{\delta}{:}\boldsymbol{d}^f)\boldsymbol{\delta} - 2\mu^f\boldsymbol{d}^f \tag{3.142}$$

which represents the *Newton's viscosity law of fluids*. A consequence of Stokes' assumption is that the average normal viscous stress is always zero, cf. [409], and the deviatoric stress tensor implies primarily viscous shear stress effects. We note that for pure fluid flow the momentum equation (3.121) with Newton's viscosity law (3.142) is commonly referred to as the *Navier-Stokes equation*.

For further needs the divergence of the deviatoric stress tensor (3.142) gives

$$\nabla\cdot(\varepsilon_f\boldsymbol{\tau}^f) = \tfrac{2}{3}\nabla(\varepsilon_f\mu^f\nabla\cdot\boldsymbol{v}^f) - 2\nabla\cdot(\mu^f\varepsilon_f\boldsymbol{d}^f) \tag{3.143}$$

Since[5]

$$\begin{aligned}\nabla\cdot(\varepsilon_f\boldsymbol{d}^f) &= \tfrac{1}{2}\nabla^2(\varepsilon_f\boldsymbol{v}^f) + \tfrac{1}{2}\varepsilon_f\nabla(\nabla\cdot\boldsymbol{v}^f)\\ &\quad -\tfrac{1}{2}\boldsymbol{v}^f\nabla^2\varepsilon_f - \tfrac{1}{2}\big(\nabla\boldsymbol{v}^f - (\nabla\boldsymbol{v}^f)^T\big)\cdot\nabla\varepsilon_f\end{aligned} \tag{3.144}$$

we can simplify (3.143) to

$$\nabla\cdot(\varepsilon_f\boldsymbol{\tau}^f) = -\mu^f\nabla^2(\varepsilon_f\boldsymbol{v}^f) \tag{3.145}$$

under the specific assumptions:

- The spatial variability of the fluid viscosity is negligible, i.e., $\|\varepsilon_f\boldsymbol{d}^f\cdot\nabla\mu^f\| \approx 0$.
- Applied to the stress tensor the vector field $\boldsymbol{v}^f$ is considered solenoidal (2.84) having $\nabla\cdot\boldsymbol{v}^f = \boldsymbol{\delta}{:}\boldsymbol{d}^f = 0$, which corresponds to the assumption of incompressibility usually made in classic fluid mechanics.
- For the stress tensor the second derivative of volume fraction is negligible $\nabla^2\varepsilon_f \approx 0$ and the antisymmetric rate of deformation tensor associated with the gradient of volume fraction vanishes: $\big(\nabla\boldsymbol{v}^f - (\nabla\boldsymbol{v}^f)^T\big)\cdot\nabla\varepsilon_f \approx \boldsymbol{0}$.

[5] In index notation we derive (dropping phase indices for the sake of simplicity)

$$\begin{aligned}\tfrac{\partial}{\partial x_j}\big[\varepsilon\tfrac{1}{2}(\tfrac{\partial v_i}{\partial x_j} + \tfrac{\partial v_j}{\partial x_i})\big] &= \tfrac{1}{2}\tfrac{\partial}{\partial x_j}\big[\tfrac{\partial(\varepsilon v_i)}{\partial x_j} + \tfrac{\partial(\varepsilon v_j)}{\partial x_i} - v_i\tfrac{\partial\varepsilon}{\partial x_j} - v_j\tfrac{\partial\varepsilon}{\partial x_i}\big]\\ &= \tfrac{1}{2}\tfrac{\partial^2(\varepsilon v_i)}{\partial x_j\partial x_j} + \tfrac{1}{2}\tfrac{\partial^2(\varepsilon v_j)}{\partial x_i\partial x_j} - \tfrac{1}{2}\tfrac{\partial}{\partial x_j}\big(v_i\tfrac{\partial\varepsilon}{\partial x_j}\big) - \tfrac{1}{2}\tfrac{\partial}{\partial x_j}\big(v_j\tfrac{\partial\varepsilon}{\partial x_i}\big)\\ &= \tfrac{1}{2}\tfrac{\partial^2(\varepsilon v_i)}{\partial x_j\partial x_j} + \tfrac{1}{2}\varepsilon\tfrac{\partial^2 v_j}{\partial x_i\partial x_j} - \tfrac{1}{2}v_i\tfrac{\partial^2\varepsilon}{\partial x_j\partial x_j} - \tfrac{1}{2}\big(\tfrac{\partial v_i}{\partial v_j} - \tfrac{\partial v_j}{\partial v_i}\big)\tfrac{\partial\varepsilon}{\partial x_j}\end{aligned}$$

The following restriction on the dynamic viscosity μ^f results from the entropy inequality (3.125):

$$\mu^f \geq 0 \tag{3.146}$$

The dynamic viscosity μ^f can be considered as a thermodynamic function

$$\mu^f = \mu^f(\omega_k^f, T^f) \tag{3.147}$$

which will be described as EOS further below.

3.8.5.2 Deviatoric Drag of Fluid Momentum Exchange $\boldsymbol{f}_\tau^f$

The truncated Taylor series expansion of $\boldsymbol{f}_\tau^f = \boldsymbol{f}_\tau^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^f, \nabla T^\alpha)$ for 1st order, however, extended by terms in $\boldsymbol{v}^{fs}$ up to third order gives

$$\begin{aligned}\boldsymbol{f}_\tau^f = &-\frac{\partial \boldsymbol{f}_\tau^f}{\partial \boldsymbol{v}^{fs}}\bigg|_0 \cdot \boldsymbol{v}^{fs} - \frac{\partial \boldsymbol{f}_\tau^f}{\partial \boldsymbol{d}^f}\bigg|_0 \cdot \boldsymbol{d}^f - \sum_k^{N^f} \frac{\partial \boldsymbol{f}_\tau^f}{\partial \nabla\omega_k^f}\bigg|_0 \cdot \nabla\omega_k^f - \frac{\partial \boldsymbol{f}_\tau^f}{\partial \nabla T^f}\bigg|_0 \cdot \nabla T^f \\ &-\tfrac{1}{2}\frac{\partial^2 \boldsymbol{f}_\tau^f}{\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\boldsymbol{v}^{fs}) - \tfrac{1}{6}\frac{\partial^3 \boldsymbol{f}_\tau^f}{\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\boldsymbol{v}^{fs}\boldsymbol{v}^{fs})\end{aligned} \tag{3.148}$$

where other HOT are neglected and a negative sign is used due to the entropy restriction $\boldsymbol{f}_\tau^f \leq 0$ in (3.125). We define for the second-order tensors:

$$\begin{aligned}\boldsymbol{R}_1^f &= \frac{\partial \boldsymbol{f}_\tau^f}{\partial \boldsymbol{v}^{fs}}\bigg|_0 \\ \boldsymbol{A}_k^f &= \frac{\partial \boldsymbol{f}_\tau^f}{\partial \nabla\omega_k^f}\bigg|_0 \\ \boldsymbol{B}^f &= \frac{\partial \boldsymbol{f}_\tau^f}{\partial \nabla T^f}\bigg|_0\end{aligned} \tag{3.149}$$

Considering isotropic conditions we find for the 3rd-order tensors:

$$\frac{\partial \boldsymbol{f}_\tau^f}{\partial \boldsymbol{d}^f}\bigg|_0 = \tfrac{1}{2}\frac{\partial^2 \boldsymbol{f}_\tau^f}{\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}}\bigg|_0 = \boldsymbol{0} \tag{3.150}$$

For the 4th-order symmetric tensor we prefer the following approximation:

$$\tfrac{1}{6}\frac{\partial^3 \boldsymbol{f}_\tau^f}{\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}}\bigg|_0 \approx \boldsymbol{R}_2^f \Im_F \|\boldsymbol{v}^{fs}\| \boldsymbol{\delta} \tag{3.151}$$

where $\boldsymbol{R}_2^f$ is a second-order tensor and $\Im_F$ is an inertial coefficient. Using (3.149), (3.150), and (3.151) the drag of momentum exchange (3.148) reads

$$\boldsymbol{f}_\tau^f = -\big(\boldsymbol{R}_1^f \cdot \boldsymbol{v}^{fs} + \boldsymbol{R}_2^f \Im_F \|\boldsymbol{v}^{fs}\| \cdot \boldsymbol{v}^{fs} + \sum_k^{N^f} \boldsymbol{A}_k^f \cdot \nabla \omega_k^f + \boldsymbol{B}^f \cdot \nabla T^f\big) \tag{3.152}$$

The third and fourth term of the RHS of (3.152) represent actions on the drag of momentum exchange which are controlled by the mass fraction and temperature gradients. Commonly, those *cross effects* on the momentum exchange are very small. For the sake of simplicity we shall assume:

- Dependency of mass fraction and temperature gradients on the drag of momentum exchange in form of cross effects are negligible, i.e., $\boldsymbol{A}_k^f \approx \boldsymbol{0}$, $\boldsymbol{B}^f \approx \boldsymbol{0}$.

Typically in porous-media problems the drag parameter for the fluid momentum exchange is related to the dynamic viscosity of the fluid μ^f and the *permeability* of the fluid phase, which is defined by

$$\boldsymbol{k}^f = \varepsilon_f^2 \mu^f (\boldsymbol{R}_1^f)^{-1} = \big(\varepsilon_f^2 \rho^f \mu^f (\boldsymbol{R}_2^f)^{-1}\big)^2 \tag{3.153}$$

where $\boldsymbol{k}^f$ is the *intrinsic permeability tensor* of the $f-$phase. With (3.153) the final form of the drag of momentum exchange of the fluid phase reads

$$\boldsymbol{f}_\tau^f = -\varepsilon_f^2 \underbrace{\mu^f (\boldsymbol{k}^f)^{-1}}_{\text{Darcy}} \cdot \boldsymbol{v}^{fs} - \varepsilon_f^2 \underbrace{\rho^f \mu^f (\boldsymbol{k}^f)^{-1/2} \Im_F \|\boldsymbol{v}^{fs}\|}_{\text{Forchheimer}} \cdot \boldsymbol{v}^{fs} \tag{3.154}$$

where the first term of the RHS of (3.154) describes the *Darcy flow* effect [392] and the second term is recognized as *Forchheimer flow* effect on the momentum drag of porous-media flow in which $\Im_F$ represent the Forchheimer coefficient, where there are different formulations and derivations in the literature, cf. [296, 534, 562]. Nield and Bejan [389] use a dimensionless form-drag constant c_F, which is related to the Forchheimer coefficient $\Im_F$ as

$$\Im_F = \frac{\varepsilon_f}{\mu^f} c_F \tag{3.155}$$

The above introduced material parameters for the fluid momentum exchange $\boldsymbol{f}_\tau^f$ are restricted by the entropy inequality (3.125) according to

$$\|\boldsymbol{k}^f\| > 0 \quad \mu^f \geq 0 \quad \Im_F \geq 0 \tag{3.156}$$

3.8.5.3 Solid Stress Tensor $\boldsymbol{\sigma}^s$

The relation between the solid stress tensor $\boldsymbol{\sigma}^s$ and the solid-phase free energy A^s given by

$$\boldsymbol{\sigma}^s = -\rho^s \frac{\partial A^s}{\partial \boldsymbol{\epsilon}^s} + p^s \boldsymbol{\delta} \tag{3.157}$$

results from the Coleman and Noll method's evaluation (3.107), where p^s is the thermodynamic pressure of the solid phase s. Furthermore, the dependence of A^s (and accordingly $\boldsymbol{\sigma}^s$ and p^s) is restricted by

$$A^s = A^s(\varepsilon_f, \rho^s, \boldsymbol{\epsilon}^s, \omega_k^s, T^s) \tag{3.158}$$

The term $\rho^s \partial A^s / \partial \boldsymbol{\epsilon}^s$ in (3.157) can be identified as the non-equilibrium solid stress

$$\boldsymbol{\tau}^s = \rho^s \frac{\partial A^s}{\partial \boldsymbol{\epsilon}^s} \tag{3.159}$$

At the thermodynamic equilibrium we have to require

$$\boldsymbol{\tau}^s \Big|_e = \mathbf{0} \tag{3.160}$$

To satisfy this equilibrium constraint the non-equilibrium stress of the solid phase $\boldsymbol{\tau}^s$ must be independent of $\varepsilon_f, \rho^s, \omega_k^s$ and T^s, i.e.,

$$\boldsymbol{\tau}^s = \boldsymbol{\tau}^s(\boldsymbol{\epsilon}^s) \tag{3.161}$$

and the truncated Taylor series expansion of $\boldsymbol{\tau}^s$ yields

$$\boldsymbol{\tau}^s = \frac{\partial \boldsymbol{\tau}^s}{\partial \boldsymbol{\epsilon}^s}\Big|_0 \cdot \boldsymbol{\epsilon}^s \tag{3.162}$$

In (3.162) a 4th-order deviatoric stress tensor appears

$$\boldsymbol{t}^s = \frac{\partial \boldsymbol{\tau}^s}{\partial \boldsymbol{\epsilon}^s}\Big|_0 \tag{3.163}$$

which can be simplified if we assume

- The solid phase s is isotropic. The deviatoric stress tensor $\boldsymbol{t}^s$ is symmetric. The solid phase can be considered as an elastic material.

Then

$$\begin{aligned} \boldsymbol{\tau}^s &= \boldsymbol{t}^s \cdot \boldsymbol{\epsilon}^s = \lambda^s(\boldsymbol{\delta}:\boldsymbol{\epsilon}^s)\boldsymbol{\delta} + 2\mu^s \boldsymbol{\epsilon}^s \\ \boldsymbol{\sigma}^s &= p^s \boldsymbol{\delta} - \boldsymbol{\tau}^s \end{aligned} \tag{3.164}$$

where λ^s and μ^s are the Lamé constants. The constitutive expression for $\boldsymbol{\tau}^s$ in (3.164) represents the *Hook's law* for isotropic linear-elastic continua. The elastic material constants λ^s and μ^s are usually expressed by the shear modulus G, Young's (or elastic) modulus E and Poisson's ratio ν as follows:

$$\begin{aligned}\lambda^s &= \frac{E\nu}{(1+\nu)(1-2\nu)} = \frac{2G\nu}{1-2\nu}\\ \mu^s &= G = \frac{E}{2(1+\nu)}\end{aligned} \tag{3.165}$$

Since strain $\boldsymbol{\epsilon}^s$ and displacement $\boldsymbol{u}^s$ are related according to (3.97), the deviatoric stress tensor $\boldsymbol{\tau}^s$ can be expressed as

$$\boldsymbol{\tau}^s = \boldsymbol{t}^s \cdot \boldsymbol{\epsilon}^s = \boldsymbol{t}^s \cdot (\boldsymbol{L} \cdot \boldsymbol{u}^s) \tag{3.166}$$

with the *elasticity matrix*

$$\begin{aligned}\boldsymbol{t}^s &= \begin{pmatrix} \lambda^s+2G & \lambda^s & \lambda^s & 0 & 0 & 0\\ \lambda^s & \lambda^s+2G & \lambda^s & 0 & 0 & 0\\ \lambda^s & \lambda^s & \lambda^s+2G & 0 & 0 & 0\\ 0 & 0 & 0 & G & 0 & 0\\ 0 & 0 & 0 & 0 & G & 0\\ 0 & 0 & 0 & 0 & 0 & G \end{pmatrix}\\ &= \tfrac{E}{(1+\nu)(1-2\nu)} \begin{pmatrix} 1-\nu & \nu & \nu & 0 & 0 & 0\\ \nu & 1-\nu & \nu & 0 & 0 & 0\\ \nu & \nu & 1-\nu & 0 & 0 & 0\\ 0 & 0 & 0 & (1-2\nu)/2 & 0 & 0\\ 0 & 0 & 0 & 0 & (1-2\nu)/2 & 0\\ 0 & 0 & 0 & 0 & 0 & (1-2\nu)/2 \end{pmatrix}\end{aligned} \tag{3.167}$$

For the material coefficients the thermodynamic restrictions require

$$\lambda^s \geq 0 \qquad \mu^s = G \geq 0 \qquad E \geq 0 \qquad 0 \leq \nu \leq \tfrac{1}{2} \tag{3.168}$$

where with $\nu = \frac{1}{2}$ the solid material is incompressible.

3.8.5.4 Heat Flux Vector $\boldsymbol{j}_T^\alpha$

The Taylor series expansion for the heat flux vector $\boldsymbol{j}_T^\alpha = \boldsymbol{j}_T^\alpha(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^\alpha, \nabla T^\alpha)$ of the $\alpha-$phase ($\alpha = f, s$) up to third order for ∇T^α product terms becomes

$$\begin{aligned}\boldsymbol{j}_T^\alpha = &-\frac{\partial \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{fs}}\bigg|_0 \cdot \boldsymbol{v}^{fs} - \frac{\partial \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{d}^f}\bigg|_0 \cdot \boldsymbol{d}^f - \sum_k^{N^\alpha} \frac{\partial \boldsymbol{j}_T^\alpha}{\partial \nabla\omega_k^\alpha}\bigg|_0 \cdot \nabla\omega_k^\alpha - \frac{\partial \boldsymbol{j}_T^\alpha}{\partial \nabla T^\alpha}\bigg|_0 \cdot \nabla T^\alpha\\ &-\tfrac{1}{2}\frac{\partial^2 \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{fs}\partial\nabla T^\alpha}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\nabla T^\alpha) - \tfrac{1}{2}\frac{\partial^2 \boldsymbol{j}_T^\alpha}{\partial\nabla T^\alpha\partial\nabla T^\alpha}\bigg|_0 \cdot (\nabla T^\alpha \nabla T^\alpha)\\ &-\tfrac{1}{6}\frac{\partial^3 \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}\partial\nabla T^\alpha}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\boldsymbol{v}^{fs}\nabla T^\alpha) - \tfrac{1}{6}\frac{\partial^3 \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{fs}\partial\nabla T^\alpha\partial\nabla T^\alpha}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\nabla T^\alpha\nabla T^\alpha)\\ &-\tfrac{1}{6}\frac{\partial^3 \boldsymbol{j}_T^\alpha}{\partial\nabla T^\alpha\partial\nabla T^\alpha\partial\nabla T^\alpha}\bigg|_0 \cdot (\nabla T^\alpha\nabla T^\alpha\nabla T^\alpha)\end{aligned} \tag{3.169}$$

where we again have assumed that the material coefficients of the $\alpha-$phase depend only on the properties of that phase. In (3.169) a negative sign is used due to the entropy restriction $\boldsymbol{j}_T^\alpha \leq 0$ in (3.125). For an isotropic medium the odd-order tensorial quantities vanish in (3.169). Furthermore, it is assumed that only first-order approximation with respect to ∇T^α is considered. Using the following definitions for the remaining second and fourth tensors in (3.169) as

$$
\begin{aligned}
\boldsymbol{U}_T^\alpha &= \left.\frac{\partial \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{fs}}\right|_0 = U_T^\alpha \boldsymbol{\delta} = U_T^\alpha \delta_{ij} \\
\boldsymbol{N}_k^\alpha &= \left.\frac{\partial \boldsymbol{j}_T^\alpha}{\partial \nabla \omega_k^f}\right|_0 = N_k^\alpha \boldsymbol{\delta} = N_k^\alpha \delta_{ij} \\
\boldsymbol{\Lambda}_0^\alpha &= \left.\frac{\partial \boldsymbol{j}_T^\alpha}{\partial \nabla T^\alpha}\right|_0 = \Lambda^\alpha \boldsymbol{\delta} = \Lambda^\alpha \delta_{ij} \\
\boldsymbol{\Lambda}_1^\alpha &= \tfrac{1}{6} \left.\frac{\partial^3 \boldsymbol{j}_T^\alpha}{\partial \boldsymbol{v}^{fs} \partial \boldsymbol{v}^{fs} \partial \nabla T^\alpha}\right|_0 = \Lambda^\alpha_{(1)ijmn} \\
&= \bar{\alpha}_T^\alpha \delta_{ij}\delta_{mn} + \frac{\bar{\alpha}_L^\alpha - \bar{\alpha}_T^\alpha}{2} (\delta_{im}\delta_{jn} + \delta_{in}\delta_{jm})
\end{aligned} \tag{3.170}
$$

the heat flux vector $\boldsymbol{j}_T^\alpha$ becomes

$$
\boldsymbol{j}_T^\alpha = -\boldsymbol{\Lambda}^\alpha \cdot \nabla T^\alpha - U_T^\alpha \boldsymbol{v}^{fs} - \sum_k^{N^\alpha} N_k^\alpha \nabla \omega_k^\alpha \tag{3.171}
$$

in which the 2nd-order *tensor of hydrodynamic thermodispersion* is introduced as

$$
\boldsymbol{\Lambda}^\alpha = \boldsymbol{\Lambda}_0^\alpha + \boldsymbol{\Lambda}_1^\alpha \cdot (\boldsymbol{v}^{fs}\boldsymbol{v}^{fs}) = \Lambda^\alpha \boldsymbol{\delta} + \boldsymbol{\Lambda}_{\text{mech}}^\alpha \tag{3.172}
$$

consisting of two parts: (1) the *tensor of thermal conductivity* $\boldsymbol{\Lambda}_0^\alpha = \Lambda^\alpha \boldsymbol{\delta}$ and (2) the *tensor of mechanical thermodispersion* $\boldsymbol{\Lambda}_{\text{mech}}^\alpha$ given by

$$
\boldsymbol{\Lambda}_{\text{mech}}^\alpha = \bar{\alpha}_T^\alpha (\boldsymbol{v}^{fs} \cdot \boldsymbol{v}^{fs}) \boldsymbol{\delta} + (\bar{\alpha}_L^\alpha - \bar{\alpha}_T^\alpha) \boldsymbol{v}^{fs} \otimes \boldsymbol{v}^{fs} \tag{3.173}
$$

where $\bar{\alpha}_L^\alpha$ and $\bar{\alpha}_T^\alpha$ represent the specific longitudinal and transverse thermodispersivity, respectively. In contrast to the form (3.173) the *classic* dispersion models developed by Scheidegger [460] and Bear [33] postulate only a linear velocity dependence for the mechanical dispersion $\boldsymbol{\Lambda}_{\text{mech}}^\alpha$ according to

$$
\boldsymbol{\Lambda}_{\text{mech}}^\alpha = \alpha_T^\alpha \|\boldsymbol{v}^{fs}\| \boldsymbol{\delta} + (\alpha_L^\alpha - \alpha_T^\alpha) \frac{\boldsymbol{v}^{fs} \otimes \boldsymbol{v}^{fs}}{\|\boldsymbol{v}^{fs}\|} \tag{3.174}
$$

where with $\alpha_L^\alpha = \bar{\alpha}_L^\alpha \|\boldsymbol{v}^{fs}\|$ and $\alpha_T^\alpha = \bar{\alpha}_T^\alpha \|\boldsymbol{v}^{fs}\|$ new longitudinal and transverse thermodispersivity coefficients appear, respectively. The Scheidegger-Bear dispersion model (3.174) is commonly used in practice. However, it is important to note that in an isotropic medium the first-order approximation of $\boldsymbol{j}_T^\alpha$ explicitly contains only $\boldsymbol{v}^{fs}\boldsymbol{v}^{fs}$ terms and not the $\boldsymbol{v}^{fs}$ terms. The material parameters for the heat flux $\boldsymbol{j}_T^\alpha$ appearing in (3.171), (3.172), and (3.174) are restricted by the entropy inequality (3.125) according to

$$\Lambda^\alpha \geq 0 \quad \alpha_L^\alpha \geq 0 \quad \alpha_T^\alpha \geq 0 \quad U_T^\alpha \geq 0 \quad N_k^\alpha \geq 0 \tag{3.175}$$

In (3.171) the heat flux is also affected by cross effects driven by the flow velocity $\boldsymbol{v}^{fs}$ and the mass fraction gradient $\nabla \omega_k^\alpha$ of species k. The influence of the concentration (mass) gradient on the heat flux is known as *Dufour effect*, where N_k^α corresponds to the Dufour coefficient. It is apparent that if mechanical and Dufour effects are neglected, we recover the conventional form of the heat flux as

$$\boldsymbol{j}_T^\alpha = -\boldsymbol{\Lambda}^\alpha \cdot \nabla T^\alpha \tag{3.176}$$

known as the *Fourier heat flux*, where the tensor of hydrodynamic thermodispersion $\boldsymbol{\Lambda}^\alpha$ is used in the form

$$\begin{aligned}\boldsymbol{\Lambda}^\alpha &= \boldsymbol{\Lambda}_0^\alpha + \boldsymbol{\Lambda}_{\text{mech}}^\alpha \\ &= (\Lambda^\alpha + \alpha_T^\alpha \|\boldsymbol{v}^{fs}\|)\boldsymbol{\delta} + (\alpha_L^\alpha - \alpha_T^\alpha)\frac{\boldsymbol{v}^{fs} \otimes \boldsymbol{v}^{fs}}{\|\boldsymbol{v}^{fs}\|}\end{aligned} \tag{3.177}$$

3.8.5.5 Species Mass Flux Vector $\boldsymbol{j}_k^f$

Similarly to the heat flux, the mass flux vector $\boldsymbol{j}_k^f = \boldsymbol{j}_k^f(\boldsymbol{v}^{fs}, \boldsymbol{d}^f, \nabla\omega_k^f, \nabla T^\alpha)$ of the species $(k = 1, \ldots, N^f)$ in the fluid phase $(f = l, g)$ is developed via a Taylor series expansion up to third order now for $\nabla\omega_k^f$ product terms. It yields

$$\begin{aligned}\boldsymbol{j}_k^f = &-\frac{\partial \boldsymbol{j}_k^f}{\partial \boldsymbol{v}^{fs}}\bigg|_0 \cdot \boldsymbol{v}^{fs} - \frac{\partial \boldsymbol{j}_k^f}{\partial \boldsymbol{d}^f}\bigg|_0 \cdot \boldsymbol{d}^f - \frac{\partial \boldsymbol{j}_k^f}{\partial \nabla\omega_k^f}\bigg|_0 \cdot \nabla\omega_k^f - \frac{\partial \boldsymbol{j}_k^f}{\partial \nabla T^f}\bigg|_0 \cdot \nabla T^f \\ &-\tfrac{1}{2}\frac{\partial^2 \boldsymbol{j}_k^f}{\partial \boldsymbol{v}^{fs}\partial\nabla\omega_k^f}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\nabla\omega_k^f) - \tfrac{1}{2}\frac{\partial^2 \boldsymbol{j}_k^f}{\partial\nabla\omega_k^f\partial\nabla\omega_k^f}\bigg|_0 \cdot (\nabla\omega_k^f\nabla\omega_k^f) \\ &-\tfrac{1}{6}\frac{\partial^3 \boldsymbol{j}_k^f}{\partial \boldsymbol{v}^{fs}\partial \boldsymbol{v}^{fs}\partial\nabla\omega_k^f}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\boldsymbol{v}^{fs}\nabla\omega_k^f) - \tfrac{1}{6}\frac{\partial^3 \boldsymbol{j}_k^f}{\partial \boldsymbol{v}^{fs}\partial\nabla\omega_k^f\partial\nabla\omega_k^f}\bigg|_0 \cdot (\boldsymbol{v}^{fs}\nabla\omega_k^f\nabla\omega_k^f) \\ &-\tfrac{1}{6}\frac{\partial^3 \boldsymbol{j}_k^f}{\partial\nabla\omega_k^f\partial\nabla\omega_k^f\partial\nabla\omega_k^f}\bigg|_0 \cdot (\nabla\omega_k^f\nabla\omega_k^f\nabla\omega_k^f)\end{aligned} \tag{3.178}$$

where a negative sign is used due to the entropy restriction $\boldsymbol{j}_k^f \leq 0$ in (3.125). In a direct analogy to the heat flux we assume that the medium is isotropic and that only a first-order approximation with respect to $\nabla\omega_k^f$ is considered. Finally, we find for the species mass flux $\boldsymbol{j}_k^f$ the following expression:

$$\boldsymbol{j}_k^f = -\rho^f \boldsymbol{D}_k^f \cdot \nabla\omega_k^f - U_C^f \boldsymbol{v}^{fs} - \boldsymbol{M}^f \nabla T^f \tag{3.179}$$

with the 2nd-order *tensor of hydrodynamic dispersion* $\boldsymbol{D}_k^f$ of species k

$$\boldsymbol{D}_k^f = \boldsymbol{D}_{k,0}^f + D_{\text{mech}}^f \tag{3.180}$$

consisting of the *tensor of diffusion*

$$\boldsymbol{D}_{k,0}^f = D_k^f \boldsymbol{\delta} \tag{3.181}$$

where D_k^f is the coefficient of molecular diffusion of species k of the fluid phase f in the porous medium, and the *tensor of mechanical dispersion*[6] of the porous medium

$$\boldsymbol{D}_{\text{mech}}^f = \beta_T^f \|\boldsymbol{v}^{fs}\| \boldsymbol{\delta} + (\beta_L^f - \beta_T^f) \frac{\boldsymbol{v}^{fs} \otimes \boldsymbol{v}^{fs}}{\|\boldsymbol{v}^{fs}\|} \tag{3.182}$$

written for the Scheidegger-Bear dispersion model, where β_L^f and β_T^f are the longitudinal and transverse dispersivities, respectively. In (3.179) cross effects for the mass flux are incorporated due to $\boldsymbol{v}^{fs}$ and ∇T^f. The temperature influence is known as *Soret effect* (or thermodiffusion), where $\boldsymbol{M}^f$ describes the Soret coefficient. Mechanical and Soret effects are commonly negligible and the species mass flux (3.179) reduces to the well-know linear *Fick's law* of macroscopic hydrodynamic dispersion

$$\boldsymbol{j}_k^f = -\rho^f \boldsymbol{D}_k^f \cdot \nabla\omega_k^f \tag{3.183}$$

[6] In 3D Cartesian coordinates, with v_1, v_2 and v_3 denoting the velocity components in the x_1, x_2 and x_3 directions, respectively, and $v = \|\boldsymbol{v}^{fs}\|$, we obtain from (3.182), dropping phase indices for convenience

$$\begin{aligned}
D_{\text{mech},11} &= \beta_T v + (\beta_L - \beta_T)\tfrac{v_1^2}{v} = \tfrac{1}{v}\big(\beta_L v_1^2 + \beta_T v_2^2 + \beta_T v_3^2\big)\\
D_{\text{mech},22} &= \beta_T v + (\beta_L - \beta_T)\tfrac{v_2^2}{v} = \tfrac{1}{v}\big(\beta_T v_1^2 + \beta_L v_2^2 + \beta_T v_3^2\big)\\
D_{\text{mech},33} &= \beta_T v + (\beta_L - \beta_T)\tfrac{v_3^2}{v} = \tfrac{1}{v}\big(\beta_T v_1^2 + \beta_T v_2^2 + \beta_L v_3^2\big)\\
D_{\text{mech},12} &= (\beta_L - \beta_T)\tfrac{v_1 v_2}{v} = D_{\text{mech},21}\\
D_{\text{mech},13} &= (\beta_L - \beta_T)\tfrac{v_1 v_3}{v} = D_{\text{mech},31}\\
D_{\text{mech},23} &= (\beta_L - \beta_T)\tfrac{v_2 v_3}{v} = D_{\text{mech},32}
\end{aligned}$$

where the tensor of hydrodynamic dispersion $\boldsymbol{D}_k^f$ is used in the form

$$\begin{aligned}\boldsymbol{D}_k^f &= \boldsymbol{D}_{k,0}^f + \boldsymbol{D}_{\text{mech}}^f \\ &= (D_k^f + \beta_T^f \|\boldsymbol{v}^{fs}\|)\boldsymbol{\delta} + (\beta_L^f - \beta_T^f)\frac{\boldsymbol{v}^{fs} \otimes \boldsymbol{v}^{fs}}{\|\boldsymbol{v}^{fs}\|}\end{aligned} \tag{3.184}$$

We find for the hydrodynamic dispersion $\boldsymbol{D}_k^f$ that the dependency on the species k is only associated with the coefficient of molecular diffusion D_k^f. It is important to note that the molecular diffusion coefficient D_k^f of the species k in the fluid phase f of the porous medium is usually smaller than the corresponding diffusion coefficient $\breve{D}_k^f$ in an open fluid body due to geometric effects of the porous medium [37, 38]. They are related by

$$D_k^f = T_*^f \breve{D}_k^f \quad (0 \leq T_*^f \leq 1) \tag{3.185}$$

with the *tortuosity* T_*^f, which ranges between zero and unity and can be approximated as [38]

$$T_*^f \approx \frac{\varepsilon_f^{7/3}}{(1-\varepsilon_s)^2} \quad \text{for} \quad (0 \leq \varepsilon_s < 1,\ 0 \leq \varepsilon_f \leq 1) \tag{3.186}$$

For the linear Fick's law (3.183) the dispersive mass flux $\boldsymbol{j}_k^f$ of a species k is proportional to the mass fraction gradient. However, it has been shown [232, 464] that if high concentrations of solutes occur, typically arising in concentrated brine transport, nonlinear effects become important and $\boldsymbol{j}_k^f$ should be replaced by an extended nonlinear *non-Fickian dispersion law*,

$$\boldsymbol{j}_k^f(\varepsilon_f \Im_H \|\boldsymbol{j}_k^f\| + 1) = -\rho^f \boldsymbol{D}_k^f \cdot \nabla\omega_k^f \tag{3.187}$$

where $\Im_H$ represents an additional high-concentration (HC) dispersion coefficient and $\boldsymbol{D}_k^f$ is the Scheidegger-Bear dispersion tensor according to (3.184). It has been found [464] that $\Im_H$ varies inversely with the flow velocity, i.e., $\Im_H = \Im_H(\boldsymbol{v}^{fs})$.

The material parameters for the species mass flux $\boldsymbol{j}_k^f$ introduced in (3.179), (3.181), (3.182), and (3.187) are restricted by the entropy inequality (3.125) according to

$$D_k^f \geq 0 \quad \beta_L^f \geq 0 \quad \beta_T^f \geq 0 \quad U_C^f \geq 0 \quad M^f \geq 0 \quad \Im_H \geq 0 \tag{3.188}$$

3.8.5.6 Species Reaction Rate $r_k^\alpha + R_k^\alpha$

The reaction rates r_k^α and R_k^α differ between homogeneous and heterogeneous reactions of species k in the multiphase system, respectively, where r_k^α concerns intraphase reactions and R_k^α covers interphase reactions. If a species k exists in different phases the mass conservation has to be related to the overall (summed) balance of mass, (3.119) plus (3.120), and a bulk reaction rate of species k for the multiphase system has to be taken into account. This bulk reaction rate can be defined as

$$R_k = \sum_\alpha \varepsilon_\alpha (r_k^\alpha + R_k^\alpha) = \sum_f \varepsilon_f (r_k^f + R_k^f) + \varepsilon_s (r_k^s + R_k^s) \tag{3.189}$$

For the constitutive representations of the rates the following functionals hold

$$\begin{aligned} r_k^\alpha &= r_k^\alpha(\omega_k^\alpha, T^\alpha) \\ R_k^\alpha &= R_k^\alpha(\omega_k^\alpha, T^\alpha) \\ R_k &= R_k(\omega_k^\alpha, T^\alpha) \end{aligned} \tag{3.190}$$

where the dependency on ρ^f can be discarded from (3.129) since the knowledge of $\omega_k^f \subset \omega_k^\alpha$ provides the fluid density according to (2.117) and (2.123). Since r_k^α, R_k^α and R_k possess the same functional structure, a polynomial representation of (3.190), exemplified for R_k, may be written as

$$R_k = b_k^0 (\omega^\alpha)^{n_k} + \sum_{m=1}^N b_m^1 (\omega_m^\alpha)^{n_m} + \sum_{m,n}^N b_m^2 (\omega_m^\alpha)^{n_m} (\omega_n^\alpha)^{n_n} + \ldots + b_m^N \prod_{m=1}^N (\omega_m^\alpha)^{n_m} \tag{3.191}$$

with

$$\prod_{m=1}^N (\omega_m^\alpha)^{n_m} = (\omega_1^\alpha)^{n_1} (\omega_2^\alpha)^{n_2} \ldots (\omega_N^\alpha)^{n_N} \tag{3.192}$$

where $n_k \geq 0$ corresponds to an exponent of species k and the coefficients b_k^p $(p = 0, 1, \ldots, N)$ are rate constants of species k depending on the overall reaction mechanism, which can be dependent on the temperature T^α

$$b_k^p = b_k^p(T^\alpha) \quad (p = 0, 1, \ldots, N) \tag{3.193}$$

applicable to a nonisothermal reaction mechanism. There are many reactive systems which can be broadly classified into simple and complex *kinetic reactions*. According to the mechanism of a reaction, the functional form of r_k^α, R_k^α or R_k can be very complicated and may not be representable as a polynomial in a form of (3.191) for all cases. Reaction mechanisms for irreversible (kinetic) and reversible (equilibrium) reactions will be discussed in more detail in Chap. 5.

3.8.6 *Equations of State (EOS)*

3.8.6.1 Fluid Density ρ^f

The fluid density ρ^f is composed of N^f miscible chemical species k with a partial fluid density $\rho_k^f = C_k^f = \rho^f \omega_k^f$ (mass of species k per unit volume of fluid), cf. (2.117) and (2.123), so that

$$\sum_{k=1}^{N^f} \omega_k^f = 1 \quad \text{and} \quad \rho^f = \sum_{k=1}^{N^f} C_k^f \tag{3.194}$$

holding for a mixture, where ω_k^f (and C_k^f) stands for all species present in the fluid phase f. However, it is important to note that only $N^f - 1$ of the mass fractions ω_k^f can be specified independently because the sum of the mass fractions must be unity. Let us for convenience designate the N^fth species as the one that is dependent, the constitutive relation is:

$$\omega_{N^f}^f = 1 - \sum_{k=1}^{N^f-1} \omega_k^f \tag{3.195}$$

It simply states that if we know the mass fractions of species 1 through $N^f - 1$, we know the mass fraction of species N^f. A typical example refers to a diluted aqueous phase, where water (species $k := N^f = H_2O$) is referred to as a *solvent* $\omega_{N^f}^f$ because it is the predominant species in a liquid phase, while the $N^f - 1$ species as *solutes* constitute only a small portion of the phase. In this context we define the special case of a *single-species solute*, where only one dissolved component exists and the aqueous phase is composed of two miscible species (one solute and one solvent), i.e., $N^f = 2$.

The density ρ^f is regarded as a dependent thermodynamic variable for which the following constitutive relationship, or EOS, (3.128) holds

$$\rho^f = \rho^f(p^f, \omega_k^f, T^f) \quad (k = 1, \ldots, N^f - 1) \tag{3.196}$$

It is to be noted that dependence is indicated on only $N^f - 1$ of the species mass fractions as shown in (3.195). We can differentiate (3.196) to obtain:

$$\begin{aligned} d\rho^f &= \frac{\partial \rho^f}{\partial p^f}\Big|_{\omega_k^f, T^f} dp^f + \sum_{k=1}^{N^f-1} \frac{\partial \rho^f}{\partial \omega_k^f}\Big|_{p^f, T^f} d\omega_k^f + \frac{\partial \rho^f}{\partial T^f}\Big|_{p^f, \omega_k^f} dT^f \\ &= \underbrace{\Big(\frac{1}{\rho^f}\frac{\partial \rho^f}{\partial p^f}\Big)}_{\gamma^f} \rho^f dp^f + \sum_{k=1}^{N^f-1} \underbrace{\Big(\frac{1}{\rho^f}\frac{\partial \rho^f}{\partial \omega_k^f}\Big)}_{\alpha_k^f} \rho^f d\omega_k^f + \underbrace{\Big(\frac{1}{\rho^f}\frac{\partial \rho^f}{\partial T^f}\Big)}_{-\beta^f} \rho^f dT^f \end{aligned} \tag{3.197}$$

where γ^f is the fluid *compressibility*, and α_k^f and β^f are the specific solutal and thermal *expansion coefficients*, respectively. A negative sign is introduced for the thermal expansion coefficient β^f to take into account that the fluid density decreases when temperature increases. Regarding γ^f and α_k^f it implies that the density ρ^f increases when the pressure p^f and/or the mass fractions ω_k^f increase, respectively. If (and only if) we assume that γ^f, α_k^f and β^f are constant, the integration of (3.197) immediately leads to the EOS for the fluid density ρ^f in the common form:

$$\rho^f = \rho_0^f e^{\gamma^f (p^f - p_0^f) + \sum_{k=1}^{N^f-1} \alpha_k^f (\omega_k^f - \omega_{k0}^f) - \beta^f (T^f - T_0^f)} \tag{3.198}$$

where suitable reference values appear for the density $\rho_0^f = \rho^f(p_0^f, \omega_{k0}^f, T_0^f)$ at reference pressure p_0^f, reference mass fraction ω_{k0}^f and reference temperature T_0^f. The EOS for the fluid density (3.198) is often linearly approximated in the form:

$$\rho^f = \rho_0^f \left[1 + \gamma^f (p^f - p_0^f) + \sum_{k=1}^{N^f-1} \alpha_k^f (\omega_k^f - \omega_{k0}^f) - \beta^f (T^f - T_0^f)\right] \tag{3.199}$$

and commonly γ^f, α_k^f and β^f are considered constant [389]. While for the most practical applications this assumption is valid for compressibility γ^f and specific solutal expansion α_k^f, a constant thermal expansion β^f may become inappropriate for geothermal applications where a larger temperature range has to be considered and thermal anomalies in ρ^f (such as the 4 °C anomaly for water) can also play a role (Fig. 3.7). For temperatures within the range from 0 to 100 °C, the thermal expansion of freshwater ($\omega_k^f = \omega_{k0}^f = 0$, $k = 1, \ldots, N^f - 1$, $p^f = p_0^f$) actually varies from $-0.68 \cdot 10^{-4}$ up to $7.5 \cdot 10^{-4}$ K^{-1}, and is zero at 4 °C [120]. To improve the relationship (3.199), a more accurate 6th-order polynomial $\rho^f = \rho^f(T^f)$ can be introduced. As shown in Appendix C a Taylor series expansion of the polynomial up to the 6th-order term results in a nonlinear *variable thermal expansion* $\beta^f = \beta^f(T^f)$, which is applied to the EOS in form of (3.199).

3.8.6.2 Internal Energy E^α

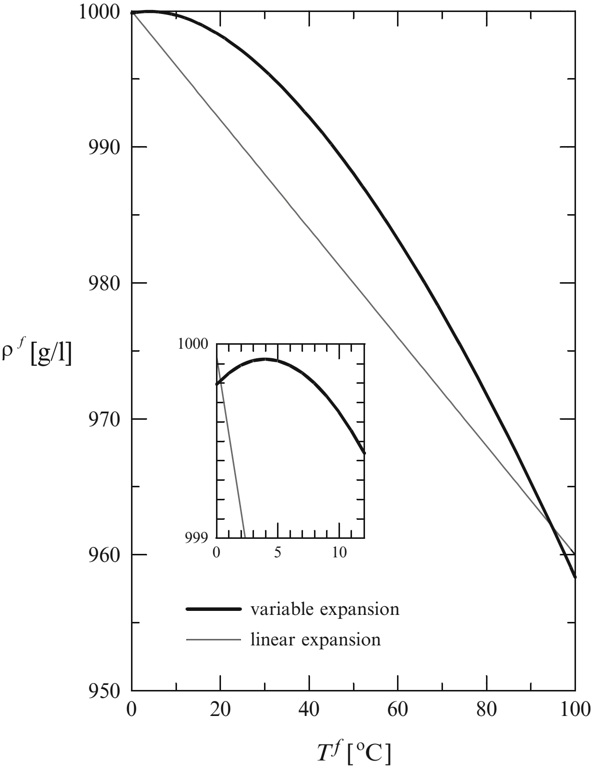
For the internal energy of the fluid phase E^f and the solid phase E^s the following dependencies exist according to (3.128):

$$\begin{aligned} E^f &= E^f(\rho^f, \omega_k^f, T^f) \\ E^s &= E^s(\varepsilon_f, \rho^s, \epsilon^s, \omega_k^s, T^s) \end{aligned} \tag{3.200}$$

Using the chain rule of differentiation it follows that

$$dE^f = \frac{\partial E^f}{\partial \rho^f}\Big|_{\omega_k^f, T^f} d\rho^f + \sum_{k=1}^{N^f} \frac{\partial E^f}{\partial \omega_k^f}\Big|_{\rho^f, T^f} d\omega_k^f + \frac{\partial E^f}{\partial T^f}\Big|_{\rho^f, \omega_k^f} dT^f \tag{3.201}$$

Fig. 3.7 Density of freshwater ρ^f (at $\omega_k^f = \omega_{k0}^f$, $p^f = p_0^f$) as a function of temperature T^f in a range of 0 and 100 °C. Close-up view indicates density anomaly at 4 °C



and

$$dE^s = \frac{\partial E^s}{\partial \rho^s}\bigg|_{\boldsymbol{\epsilon}^s, \omega_k^s, T^s} d\rho^s + \frac{\partial E^s}{\partial \boldsymbol{\epsilon}^s}\bigg|_{\rho^s, \omega_k^s, T^s} d\boldsymbol{\epsilon}^s + \sum_{k=1}^{N^s} \frac{\partial E^s}{\partial \omega_k^s}\bigg|_{\rho^s, \boldsymbol{\epsilon}^s, T^s} d\omega_k^s + \frac{\partial E^s}{\partial T^s}\bigg|_{\rho^s, \boldsymbol{\epsilon}^s, \omega_k^s} dT^s \tag{3.202}$$

where for (3.202) we have assumed that E^s depends only on properties of the solid phase s. The task here is to find

$$\frac{\partial E^\alpha}{\partial \rho^\alpha}\bigg|_{\boldsymbol{\epsilon}^s, \omega_k^\alpha, T^\alpha}, \quad \frac{\partial E^s}{\partial \boldsymbol{\epsilon}^s}\bigg|_{\rho^s, \omega_k^s, T^s}, \quad \frac{\partial E^\alpha}{\partial \omega_k^\alpha}\bigg|_{\rho^\alpha, \boldsymbol{\epsilon}^s, T^\alpha}, \quad \frac{\partial E^\alpha}{\partial T^\alpha}\bigg|_{\rho^\alpha, \boldsymbol{\epsilon}^s, \omega_k^\alpha},$$
$$(\alpha = s, f), (k = 1, \ldots, N^\alpha) \tag{3.203}$$

Taking into account from (3.67), (3.107), (3.108), (3.164), and (3.166)

$$\begin{aligned}
E^\alpha &= A^\alpha + T^\alpha S^\alpha \\
\frac{\partial A^\alpha}{\partial \omega_k^\alpha} &= \mu_k^\alpha \\
\frac{\partial A^\alpha}{\partial T^\alpha} &= -S^\alpha \qquad\qquad (\alpha = s, f) \\
\frac{\partial A^s}{\partial \boldsymbol{\epsilon}^s} &= \frac{p^s - \boldsymbol{\sigma}^s}{\rho^s} = \frac{\boldsymbol{t}^s \cdot \boldsymbol{\epsilon}^s}{\rho^s} \\
p^\alpha &= \rho^{\alpha^2} \frac{\partial A^\alpha}{\partial \rho^\alpha}
\end{aligned} \tag{3.204}$$

we obtain

$$\begin{aligned}
\left.\frac{\partial E^\alpha}{\partial \rho^\alpha}\right|_{\boldsymbol{\epsilon}^s, \omega_k^\alpha, T^\alpha} &= \frac{p^\alpha}{\rho^{\alpha^2}} - T^\alpha \frac{\partial}{\partial T^\alpha}\left(\frac{p^\alpha}{\rho^{\alpha^2}}\right)\Big|_{\rho^\alpha, \boldsymbol{\epsilon}^s, \omega_k^\alpha} \\
&= \frac{1}{\rho^{\alpha^2}}\left(p^\alpha - T^\alpha \frac{\partial p^\alpha}{\partial T^\alpha}\right) \\
\left.\frac{\partial E^s}{\partial \boldsymbol{\epsilon}^s}\right|_{\rho^s, \omega_k^s, T^s} &= \frac{1}{\rho^s}\left(\boldsymbol{t}^s \cdot \boldsymbol{\epsilon}^s - T^s \frac{\partial(\boldsymbol{t}^s \cdot \boldsymbol{\epsilon}^s)}{\partial T^s}\right) \qquad (\alpha = s, f) \\
\left.\frac{\partial E^\alpha}{\partial \omega_k^\alpha}\right|_{\rho^\alpha, \boldsymbol{\epsilon}^s, T^\alpha} &= \mu_k^\alpha - T^\alpha \frac{\partial \mu_k^\alpha}{\partial T^\alpha} \quad (k = 1, \ldots, N^\alpha) \\
\left.\frac{\partial E^\alpha}{\partial T^\alpha}\right|_{\rho^\alpha, \boldsymbol{\epsilon}^s, \omega_k^\alpha} &= c^\alpha
\end{aligned} \tag{3.205}$$

where with c^α the *specific heat capacity* of the $\alpha-$phase is introduced which is usually positive $c^\alpha > 0$. Note that c^α need not be constant. We substitute (3.205) into (3.201) and (3.202) to find the expression for the material derivatives,[7] viz.,

$$\begin{aligned}
\varepsilon_f \rho^f \frac{D^f E^f}{Dt} &= \frac{\varepsilon_f}{\rho^f}\left(p^f - T^f \frac{\partial p^f}{\partial T^f}\right)\frac{D^f \rho^f}{Dt} + \\
\varepsilon_f \rho^f \sum_k^{N^f} &\left(\mu_k^f - T^f \frac{\partial \mu_k^f}{\partial T^f}\right)\frac{D^f \omega_k^f}{Dt} + \varepsilon_f \rho^f c^f \frac{D^f T^f}{Dt}
\end{aligned} \tag{3.206}$$

[7] Using calculus manipulations the material derivative of E^f with respect to the density ρ^f can be alternatively developed for the $\frac{D^f \rho^f}{Dt}$ term:

$$\frac{\varepsilon_f}{\rho^f}\left(p^f - T^f \frac{\partial p^f}{\partial T^f}\right)\frac{D^f \rho^f}{Dt} = \frac{\varepsilon_f p^f}{\rho^f}\frac{D^f \rho^f}{Dt} + \varepsilon_f T^f \beta^f \frac{D^f p^f}{Dt}$$

where the thermal expansion coefficient (3.197), $\beta^f = -(1/\rho^f)(\partial \rho^f/\partial T^f)$, is inserted.

and

$$\varepsilon_s\rho^s\frac{D^sE^s}{Dt}=\frac{\varepsilon_s}{\rho^s}\Big(p^s-T^s\frac{\partial p^s}{\partial T^s}\Big)\frac{D^s\rho^s}{Dt}+$$
$$\Big[\varepsilon_s(\boldsymbol{t}^s\cdot\boldsymbol{\epsilon}^s)-\varepsilon_sT^s\frac{\partial(\boldsymbol{t}^s\cdot\boldsymbol{\epsilon}^s)}{\partial T^s}\Big]\frac{D^s\boldsymbol{\epsilon}^s}{Dt}+$$
$$\varepsilon_s\rho^s\sum_k^{N^s}\Big(\mu_k^s-T^s\frac{\partial\mu_k^s}{\partial T^s}\Big)\frac{D^s\omega_k^s}{Dt}+\varepsilon_s\rho^sc^s\frac{D^sT^s}{Dt} \tag{3.207}$$

In general, the *chemical potential* $\mu_k^\alpha=\mu_k^\alpha(\rho^\alpha,\omega_k^\alpha,T^\alpha)$, cf. (3.107), is a dependent variable and further constitutive relations are required. However, in most applications

- The density, solid strain and chemical effects on the internal energy are negligible,

so that E^α becomes only dependent on the temperature T^α

$$dE^\alpha=c^\alpha dT^\alpha\quad(\alpha=f,s) \tag{3.208}$$

and the material derivatives (3.206) and (3.207) simplify in

$$\varepsilon_\alpha\rho^\alpha\frac{D^\alpha E^\alpha}{Dt}=\varepsilon_\alpha\rho^\alpha c^\alpha\frac{D^\alpha T^\alpha}{Dt}\quad(\alpha=f,s) \tag{3.209}$$

If the specific heat capacity c^α is *independent* of the temperature T^α, the internal energy $E^\alpha(T^\alpha)=E^\alpha(T_0^\alpha)+\int_{T_0^\alpha}^{T^\alpha}c^\alpha dT^\alpha$ can be given explicitly

$$E^\alpha=E_0^\alpha+c^\alpha(T^\alpha-T_0^\alpha)\quad(\alpha=f,s) \tag{3.210}$$

where $E_0^\alpha=E^\alpha(T_0^\alpha)$ is a constant reference value of internal energy.

3.8.6.3 Dynamic Viscosity μ^f

The dynamic viscosity μ^f of the fluid phase $f=l,g$ is regarded as a thermodynamic function of mass fraction ω_k^f and temperature T^f, cf. (3.147):

$$\mu^f=\mu^f(\omega_k^f,T^f) \tag{3.211}$$

A truncated Taylor series expansion for μ^f around reference mass fraction ω_{k0}^f and reference temperature T_0^f up to the 3rd order gives

$$\mu^f = \mu_0^f + \frac{\partial \mu^f}{\partial T^f}\Big|_{T_0^f,\omega_{k0}^f}(T^f - T_0^f) + \tfrac{1}{2}\frac{\partial^2 \mu^f}{\partial T^{f^2}}\Big|_{T_0^f,\omega_{k0}^f}(T^f - T_0^f)^2 +$$
$$\tfrac{1}{6}\frac{\partial^3 \mu^f}{\partial T^{f^3}}\Big|_{T_0^f,\omega_{k0}^f}(T^f - T_0^f)^3 + \sum_{k=1}^{N^f}\frac{\partial \mu^f}{\partial \omega_k^f}\Big|_{T_0^f,\omega_{k0}^f}(\omega_k^f - \omega_{k0}^f) +$$
$$\sum_{k=1}^{N^f}\tfrac{1}{2}\frac{\partial^2 \mu^f}{\partial \omega_k^{f^2}}\Big|_{T_0^f,\omega_{k0}^f}(\omega_k^f - \omega_{k0}^f)^2 + \sum_{k=1}^{N^f}\tfrac{1}{6}\frac{\partial^3 \mu^f}{\partial \omega_k^{f^3}}\Big|_{T_0^f,\omega_{k0}^f}(\omega_k^f - \omega_{k0}^f)^3 \qquad (3.212)$$

where the terms $\ldots|_{T_0^f,\omega_{k0}^f}$ are constant coefficients of fluid viscosity at reference temperature and reference mass fraction, and μ_0^f is the reference fluid viscosity at reference temperature and reference mass fraction.

Furthermore, viscosity dependencies have been developed by using empirical polynomial relationships in the literature. Regarding the mass fraction dependency, particularly for high-concentration saltwater, Lever and Jackson [343] and Hassanizadeh [224] proposed the following relationship:

$$\mu^l(\omega^l) = \bar{\mu}_0^l(1 + 1.85\omega - 4.1\omega^2 + 44.5\omega^3) \qquad (3.213)$$

with

$$\begin{aligned} \omega^l &= \sum_{k=1}^{N^l-1}\omega_k^l \\ \omega &= \omega^l - \omega_0^l \quad \text{with} \quad \omega_0^l \equiv 0 \\ \bar{\mu}_0^l &= \mu^l(\omega_0^l = 0) \end{aligned} \qquad (3.214)$$

where ω^l is the overall mass fraction of the *total dissolved solids* (TDS) in the liquid (= water) phase l and $\bar{\mu}_0^l$ is a specific reference viscosity valid for $\omega_0^l = 0$ at, however, unspecified temperatures. On the other hand, an empirical relation for the temperature dependence of the dynamic viscosity μ^l of the liquid (= water) phase l has been proposed by Mercer and Pinder [371] in the form:

$$\frac{1}{\mu^l(T^l)} = \frac{1 + 0.7063\varsigma - 0.04832\varsigma^3}{\bar{\mu}_0^l} \qquad (3.215)$$

with

$$\left.\begin{aligned} \varsigma &= \frac{(T^l - T_0^l)}{100} \\ T_0^l &= 150 \end{aligned}\right\} \quad \text{for} \quad T^l \quad \text{in} \quad {}^\circ\text{C} \qquad (3.216)$$

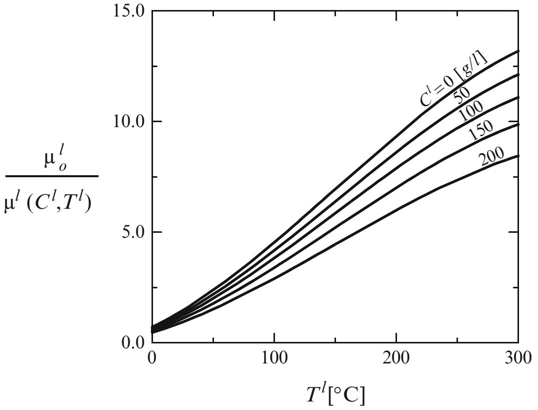


Fig. 3.8 Viscosity relation μ^l_0/μ^l of water l as function of mass concentration $C^l = \omega^l \rho^l$ [g/l] and temperature T^l [°C] using $C^l_0 = \omega^l_0 = 0$ (freshwater) and $T^l_0 = 10\,°\mathrm{C}$

where the specific reference viscosity $\bar{\mu}^l_0$ is related to a reference temperature of $150\,°\mathrm{C}$ (or $\varsigma = 0$) at, however, unspecified mass fraction of solutes. A combination of both influences yields the following expression:

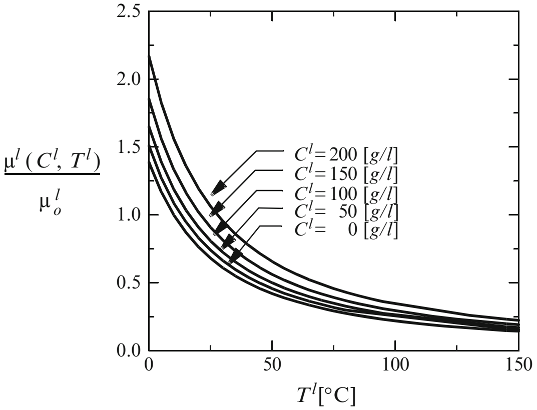
$$\frac{\bar{\mu}^l_0}{\mu^l(\omega^l, T^l)} = \frac{1 + 0.7063\varsigma - 0.04832\varsigma^3}{1 + 1.85\omega - 4.1\omega^2 + 44.5\omega^3} \tag{3.217}$$

where $\bar{\mu}^l_0 = \mu^l(0, 150°\mathrm{C})$. Employing an arbitrary reference mass fraction ω^l_0 and reference temperature T^l_0, a *viscosity relation function* f^l_μ can be obtained, which is related to these proper reference conditions:

$$\begin{aligned} f^l_\mu &= \frac{\mu^l_0}{\mu^l(\omega^l, T^l)} = \frac{\bar{\mu}^l_0}{\mu^l(\omega^l, T^l)} \frac{\mu^l(\omega^l_0, T^l_0)}{\bar{\mu}^l_0} \\ &= \frac{1+1.85\omega_{(\omega^l=\omega^l_0)}-4.1\omega^2_{(\omega^l=\omega^l_0)}+44.5\omega^3_{(\omega^l=\omega^l_0)}}{1+1.85\omega-4.1\omega^2+44.5\omega^3} \, \frac{1+0.7063\varsigma-0.04832\varsigma^3}{1+0.7063\varsigma_{(T^l=T^l_0)}-0.04832\varsigma^3_{(T^l=T^l_0)}} \end{aligned} \tag{3.218}$$

where $\mu^l_0 = \mu^l(\omega^l_0, T^l_0)$ is the reference viscosity with respect to the reference mass fraction ω^l_0 and the reference temperature T^l_0. For example, typical viscosity relations of water l are displayed in Figs. 3.8 and 3.9 for the temperature range T^l between 0 and $300\,°\mathrm{C}$ and mass concentrations $C^l = \omega^l \rho^l$ between 0 and 200 g/l for the chosen reference concentration $\omega^l_0 = C^l_0 = 0$ (freshwater) and reference temperature $T^l_0 = 10\,°\mathrm{C}$.

Fig. 3.9 Viscosity relation μ^l/μ_0^l of water l as function of mass concentration $C^l = \omega^l \rho^l$ [g/l] and temperature T^l [°C] using $C_0^l = \omega_0^l = 0$ (freshwater) and $T_0^l = 10\,°\mathrm{C}$



3.8.7 *Additional Closure Relations*

The volume fraction ε_α $(\alpha = f, s)$ (3.4) appears as a geometry-dependent variable resulting from the volume averaging process over a REV. It is convenient to express ε_α for fluid phases $f = l, g$ and the solid phase s of a porous medium as

$$\begin{aligned}\varepsilon_l &= \varepsilon s^l\\ \varepsilon_g &= \varepsilon s^g\\ \varepsilon_s &= 1-\varepsilon\end{aligned} \tag{3.219}$$

with

$$s^l + s^g = 1 \quad 0 \le s^l \le 1 \quad 0 \le s^g \le 1 \quad \varepsilon = \varepsilon_l + \varepsilon_g = 1 - \varepsilon_s \tag{3.220}$$

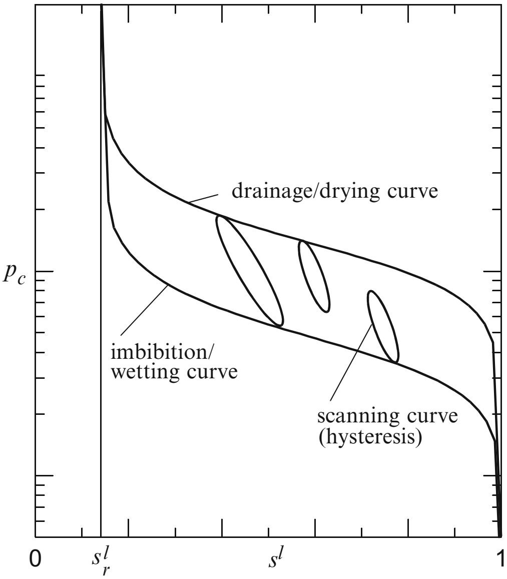
and $\sum_\alpha \varepsilon_\alpha = \varepsilon_l + \varepsilon_g + \varepsilon_s \equiv 1$ (3.5), where ε is the *porosity* (void space) and s is the *fluid saturation* referring to the dynamic liquid l and gas g phases of the porous medium. For the following considerations we assume that the liquid phase l represents the *wetting phase* and the gas phase g represents the *nonwetting phase* of the two coexisting fluids filled in the void space ε of the porous medium.

3.8.7.1 Capillary Pressure p_c

The macroscopic representation of the equilibrium with the pressure difference between adjacent nonwetting and wetting fluid phases at the interface of the two fluids in the void space of a porous medium is recorded by the macroscopic *capillary pressure* p_c, defined as

$$p_c = p^g - p^l \tag{3.221}$$

Fig. 3.10 Schematic plot of capillary pressure p_c versus liquid saturation s^l possessing hysteresis



for which constitutive relationships are required, e.g.,

$$p_c = p_c(s^l, T^l, T^g, \omega_k^l, \omega_k^g) \tag{3.222}$$

We have to note that the dependency of p_c on the liquid saturation s^l does not follow consistently from the thermodynamic dependence of the liquid pressure as stated above in (3.108). The conflict results from the difference between the pressure variable p^α of $\alpha-$phase, which is a volume average over the REV, and the capillary pressure p_c, which is basically an interface variable and accordingly should refer to as a surface average. A comprehensive discussion on this matter and an extended alternative theoretical approach can be found in Gray and Hassanizadeh [205, 206], Hassanizadeh and Gray [231], and Hassanizadeh et al. [233]. Temperature effects on capillary pressure are presented by Grant [200]. Commonly, the capillary pressure p_c is considered to be dependent on the wetted liquid phase l only, viz.,

$$p^g - p^l = p_c(s^l) \tag{3.223}$$

where numerous empirical relations exist to express $p_c(s^l)$. The explicit functional form for $p_c(s^l)$ must be considered to be specific to the combination of the pair of fluids and the porous medium, basically also dependent on the medium temperature and the chemical composition of the fluids. The function of p_c is also known to exhibit *hysteresis* in that the equilibrium value of p_c as a function of s^l is found to be dependent on the direction of the process (i.e., drainage (drying) or imbibition (wetting)). A schematic depiction of the p_c versus s^l curves is given in Fig. 3.10.

In soil science, the $p_c(s^l)$–relationship is called *retention curve* as it shows how much water is retained in a soil by the capillary pressure. Numerous empirical *parametric models* exist to describe retention curves based on fitted analytical expressions. The most common empirical relations are summarized in Appendix D. Alternatively, spline approximations of the retention curve can be useful in cases where analytical functions do not fit suitably to the experimental data, for more see also Appendix D.

3.8.7.2 Relative Permeability k_r^f

The presence of more than one fluid phase ($f = l, g$) in a porous medium has consequences on the interfacial momentum exchange $\boldsymbol{f}_\tau^f$ too. For the intrinsic permeability tensor $\boldsymbol{k}^f$ of the fluid phase f appearing in the drag term of fluid momentum exchange (3.154) a saturation-dependency is postulated in the following form:

$$\boldsymbol{k}^f = \boldsymbol{k}^f(s^f) = k_r^f(s^f)\boldsymbol{k} \quad (0 < k_r^f \leq 1) \quad (f = l, g) \tag{3.224}$$

where with k_r^f the saturation-dependent *relative permeability*, sometimes called *relative conductivity*, is introduced via a variable separation and the split $\boldsymbol{k}$ is the fluid-independent *permeability tensor* of the porous medium, which is anisotropic in general. The permeability tensor $\boldsymbol{k}$ is also termed as *saturated permeability* equivalent to the intrinsic permeability at full saturation $s^f = 1$. Various empirical relationships for $k_r^f = k_r^f(s^f)$ exist, where the most useful parametric models are summarized in Appendix D. A typical curve of $k_r^l(s^l)$ for the liquid phase l is exhibited in Fig. 3.11. Note that the relative permeability k_r^f can also imply hysteretic effects [34, 38, 422]. In case of need spline approximation for $k_r^l(s^l)$ can be beneficial to get better fits to experimental data, see Appendix D.

3.9 Complete Equations of Multiphase Flow and Transport in Deforming Porous Media

3.9.1 General Formulation

The formulation of a complete mathematical model for solving multiphase flow, mass and heat transport in deforming porous media is based on the balance laws of Sects. 3.7 and 3.8.4 in combination with the phenomenological equations of Sect. 3.8.5, the constitutive relations of Sect. 3.8.6 and the additional closure relations of Sect. 3.8.7 including the assumptions as stated in these sections. It results in a rather general set of equations as follows:

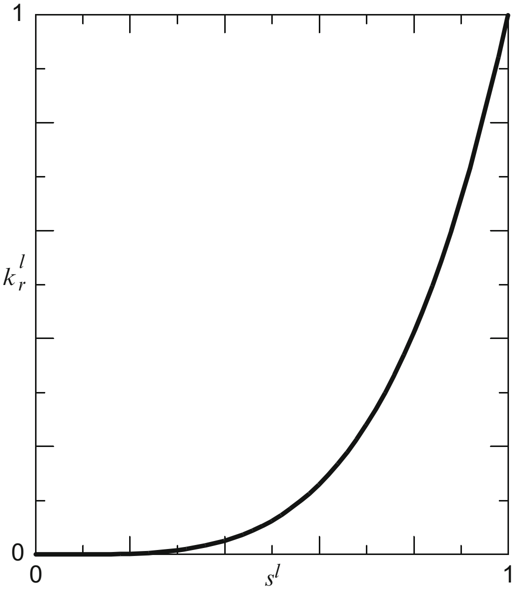


Fig. 3.11 Schematic plot of relative permeability k_r^l versus liquid saturation s^l

Mass conservation $\mathcal{M}^f$ of fluid phases $f = l, g$

$$\frac{\partial}{\partial t}(\varepsilon_f \rho^f) + \nabla \cdot (\varepsilon_f \rho^f \boldsymbol{v}^f) = \rho^f Q_f \tag{3.225}$$

Mass conservation $\mathcal{M}^s$ of solid phase s [8]

$$\frac{\partial}{\partial t}(\varepsilon_s \rho^s) + \nabla \cdot (\varepsilon_s \rho^s \boldsymbol{v}^s) = \rho^s Q_s \tag{3.226}$$

Mass conservation $\mathcal{M}_k^f$ of species k of fluid phases $f = l, g$

divergence form

$$\frac{\partial}{\partial t}(\varepsilon_f \rho^f \omega_k^f) + \nabla \cdot (\varepsilon_f \rho^f \boldsymbol{v}^f \omega_k^f) + \nabla \cdot \boldsymbol{j}_{fk} = \varepsilon_f (r_k^f + R_k^f) \tag{3.227}$$

convective form

[8] It can be alternatively expressed by introducing the relationships (3.95) and (3.100) of the solid displacement $\boldsymbol{u}^s$:

$$\frac{\partial}{\partial t}(\varepsilon_s \rho^s) + \varepsilon_s \rho^s \left(\boldsymbol{m}^T \cdot \left(\boldsymbol{L} \cdot \frac{\partial \boldsymbol{u}^s}{\partial t} \right) \right) + \nabla(\varepsilon_s \rho^s) \cdot \frac{\partial \boldsymbol{u}^s}{\partial t} = \rho^s Q_s$$

$$\varepsilon_f \rho^f \frac{\partial \omega_k^f}{\partial t} + \varepsilon_f \rho^f \boldsymbol{v}^f \cdot \nabla \omega_k^f + \nabla \cdot \boldsymbol{j}_{fk} = \varepsilon_f (r_k^f + R_k^f) - \rho^f \omega_k^f Q_f \tag{3.228}$$

Mass conservation $\mathcal{M}_k^s$ of species k of solid phase s

divergence form

$$\frac{\partial}{\partial t}(\varepsilon_s \rho^s \omega_k^s) + \nabla \cdot (\varepsilon_s \rho^s \boldsymbol{v}^s \omega_k^s) = \varepsilon_s (r_k^s + R_k^s) \tag{3.229}$$

convective form

$$\varepsilon_s \rho^s \frac{\partial \omega_k^s}{\partial t} + \varepsilon_s \rho^s \boldsymbol{v}^s \cdot \nabla \omega_k^s = \varepsilon_s (r_k^s + R_k^s) - \rho^s \omega_k^s Q_s \tag{3.230}$$

Momentum conservation $\mathcal{V}^f$ of fluid phases $f = l, g$

divergence form

$$\begin{aligned} \frac{\partial}{\partial t}(\varepsilon_f \rho^f \boldsymbol{v}^f) + \nabla \cdot (\varepsilon_f \rho^f (\boldsymbol{v}^f \boldsymbol{v}^f)) + \varepsilon_f \nabla p^f - \mu^f \nabla^2 (\varepsilon_f \boldsymbol{v}^f) &= \varepsilon_f \rho^f \boldsymbol{g} \\ -\varepsilon_f^2 \mu^f (k_r^f \boldsymbol{k})^{-1} \cdot \boldsymbol{v}^{fs} - \varepsilon_f^2 \rho^f \mu^f (k_r^f \boldsymbol{k})^{-1/2} \Im_F \|\boldsymbol{v}^{fs}\| \cdot \boldsymbol{v}^{fs} & \end{aligned} \tag{3.231}$$

convective form

$$\begin{aligned} \varepsilon_f \rho^f \frac{\partial \boldsymbol{v}^f}{\partial t} + \varepsilon_f \rho^f \boldsymbol{v}^f \cdot \nabla \boldsymbol{v}^f + \varepsilon_f \nabla p^f - \mu^f \nabla^2 (\varepsilon_f \boldsymbol{v}^f) &= \varepsilon_f \rho^f \boldsymbol{g} \\ -\varepsilon_f^2 \mu^f (k_r^f \boldsymbol{k})^{-1} \cdot \boldsymbol{v}^{fs} - \varepsilon_f^2 \rho^f \mu^f (k_r^f \boldsymbol{k})^{-1/2} \Im_F \|\boldsymbol{v}^{fs}\| \cdot \boldsymbol{v}^{fs} - \rho^f \boldsymbol{v}^f Q_f & \end{aligned} \tag{3.232}$$

Momentum conservation $\mathcal{V}^s$ of solid phase s

convective form

$$\varepsilon_s \rho^s \frac{\partial^2 \boldsymbol{u}^s}{\partial t^2} + \nabla(\varepsilon_s p^s) - \boldsymbol{L}^T \cdot \big(\varepsilon_s \boldsymbol{t}^s \cdot (\boldsymbol{L} \cdot \boldsymbol{u}^s)\big) = \varepsilon_s \rho^s \boldsymbol{g} - \rho^s \boldsymbol{v}^s Q_s \tag{3.233}$$

Energy conservation $\mathcal{E}^f + \mathcal{K}^f$ of fluid phases $f = l, g$

divergence form

$$\begin{aligned} &\frac{\partial}{\partial t}\big(\varepsilon_f \rho^f c^f (T^f - T_0^f)\big) + \nabla \cdot \big(\varepsilon_f \rho^f c^f \boldsymbol{v}^f (T^f - T_0^f)\big) \\ &-\nabla \cdot (\boldsymbol{\Lambda}_f \cdot \nabla T^f) = -\varepsilon_f p^f \nabla \cdot \boldsymbol{v}^f - \tfrac{2}{3} \varepsilon_f \mu^f (\nabla \cdot \boldsymbol{v}^f)^2 + 2 \varepsilon_f \mu^f \boldsymbol{d}^f : \boldsymbol{d}^f \\ &+\rho^f H_f \end{aligned} \tag{3.234}$$

convective form

$$\begin{aligned}\varepsilon_f\rho^f c^f\frac{\partial T^f}{\partial t}+\varepsilon_f\rho^f c^f\boldsymbol{v}^f\cdot\nabla T^f-\nabla\cdot(\boldsymbol{\Lambda}_f\cdot\nabla T^f)&=\\-\varepsilon_f p^f\nabla\cdot\boldsymbol{v}^f-\tfrac{2}{3}\varepsilon_f\mu^f(\nabla\cdot\boldsymbol{v}^f)^2+2\varepsilon_f\mu^f\boldsymbol{d}^f{:}\boldsymbol{d}^f+\rho^f H_f&\\-\rho^f c^f(T^f-T_0^f)Q_f&\end{aligned}\tag{3.235}$$

Energy conservation $\mathcal{E}^s+\mathcal{K}^s$ of solid phase s

divergence form

$$\begin{aligned}\frac{\partial}{\partial t}\big(\varepsilon_s\rho^s c^s(T^s-T_0^s)\big)+\nabla\cdot\big(\varepsilon_s\rho^s c^s\boldsymbol{v}^s(T^s-T_0^s)\big)&\\-\nabla\cdot(\boldsymbol{\Lambda}_s\cdot\nabla T^s)=-\varepsilon_s p^s\nabla\cdot\boldsymbol{v}^s+\varepsilon_s(\boldsymbol{t}^s\cdot\boldsymbol{\epsilon}^s){:}\boldsymbol{\epsilon}^s+\rho^s H_s&\end{aligned}\tag{3.236}$$

convective form

$$\begin{aligned}\varepsilon_s\rho^s c^s\frac{\partial T^s}{\partial t}+\varepsilon_s\rho^s c^s\boldsymbol{v}^s\cdot\nabla T^s-\nabla\cdot(\boldsymbol{\Lambda}_s\cdot\nabla T^s)&=\\-\varepsilon_s p^s\nabla\cdot\boldsymbol{v}^s+\varepsilon_s(\boldsymbol{t}^s\cdot\boldsymbol{\epsilon}^s){:}\boldsymbol{\epsilon}^s+\rho^s H_s-\rho^s c^s(T^s-T_0^s)Q_s&\end{aligned}\tag{3.237}$$

Constitutive relations

$$\begin{aligned}\rho^f&=\rho_0^f\big[1+\gamma^f(p^f-p_0^f)+\textstyle\sum_{k=1}^{N^f-1}\alpha_k^f(\omega_k^f-\omega_{k0}^f)\\&\quad-\beta^f(T^f)(T^f-T_0^f)\big]\\p_c&=p_c(s^l)=p^g-p^l\\k_r^f&=k_r^f(s^f)\\\boldsymbol{j}_{fk}(\Im_H\|\boldsymbol{j}_{fk}\|+1)&=-\rho^f\boldsymbol{D}_{fk}\cdot\nabla\omega_k^f\\\boldsymbol{D}_{fk}&=\varepsilon_f D_k^f\boldsymbol{\delta}+\boldsymbol{D}_{f\text{mech}}\\\boldsymbol{\Lambda}_f&=\varepsilon_f\Lambda^f\boldsymbol{\delta}+\rho^f c^f\boldsymbol{D}_{f\text{mech}}\\\boldsymbol{\Lambda}_s&=\varepsilon_s\Lambda^s\boldsymbol{\delta}\\\boldsymbol{D}_{f\text{mech}}&=\varepsilon_f\Big[\beta_T^f\|\boldsymbol{v}^{fs}\|\boldsymbol{\delta}+(\beta_L^f-\beta_T^f)\tfrac{\boldsymbol{v}^{fs}\otimes\boldsymbol{v}^{fs}}{\|\boldsymbol{v}^{fs}\|}\Big]\\\boldsymbol{t}^s&=\boldsymbol{t}^s(\lambda^s,\mu^s)\\\mu^f&=\mu^f(\omega_k^f,T^f)\end{aligned}\tag{3.238}$$

with $\boldsymbol{v}^{fs}=\boldsymbol{v}^f-\boldsymbol{v}^s$, $\boldsymbol{d}^f=\frac{1}{2}[\nabla\boldsymbol{v}^f+(\nabla\boldsymbol{v}^f)^T]$, $\varepsilon_f=\varepsilon s^f$, $\varepsilon_s=1-\varepsilon$, $\boldsymbol{v}^s=\partial\boldsymbol{u}^s/\partial t$ and $\boldsymbol{\epsilon}^s=\boldsymbol{L}\cdot\boldsymbol{u}^s$. We introduced above appropriate bulk quantities denoted by phase subscripts as follows: $\boldsymbol{j}_{fk}=\varepsilon_f\boldsymbol{j}_k^f,\boldsymbol{D}_{fk}=\varepsilon_f\boldsymbol{D}_k^f,\boldsymbol{D}_{f\text{mech}}=\varepsilon_f\boldsymbol{D}_{\text{mech}}^f,\boldsymbol{\Lambda}_\alpha=\varepsilon_\alpha\boldsymbol{\Lambda}^\alpha$. In (3.238) we made use of the fact that the mechanical dispersion is a property of the porous medium and independent of the actual transport quantity; accordingly, we substituted $\boldsymbol{\Lambda}^f_{\text{mech}} = \rho^f c^f \boldsymbol{D}^f_{\text{mech}}$. We use both divergence and convective forms of the balance statements if necessary due to mathematical reasons as discussed further below. For the energy conservation it is obvious that the convective form naturally results from replacing the internal energy by the temperture variable, cf. (3.206) and (3.207). A divergence form of energy conservation with the temperature variable results from the basic energy balance equation (3.59) by inserting (3.210). That means, their expressions (3.234) and (3.236) in terms of temperatures T^f and T^s, respectively, are possible if (and only if) the specific heat capacities c^f and c^s are assumed independent of temperatures. Such an assumption is not needed for the convective forms (3.235) and (3.237) of energy conservation.

The conservation laws (3.225)–(3.237) for the three $(l - g - s)$ phases form a closed equation system consisting of $6 + N + 3D$ equations, which can be solved for the following independent primary variables:

$$\begin{array}{l}
\left.\begin{array}{ll}
p^l & \text{from (3.225)}\\
p^g & \text{from (3.225)}\\
p^s & \text{from (3.226)}
\end{array}\right\}\ 3 \quad \text{equations}\\
\left.\begin{array}{ll}
\omega^l_k & \text{from (3.227) or (3.228)}\\
\omega^g_k & \text{from (3.227) or (3.228)}\\
\omega^s_k & \text{from (3.229) or (3.230)}
\end{array}\right\}\ N \quad \text{equations}\\
\left.\begin{array}{ll}
\boldsymbol{v}^l & \text{from (3.231) or (3.232)}\\
\boldsymbol{v}^g & \text{from (3.231) or (3.232)}\\
\boldsymbol{u}^s & \text{from (3.233)}
\end{array}\right\}\ 3D \ \text{equations}\\
\left.\begin{array}{ll}
T^l & \text{from (3.234) or (3.235)}\\
T^g & \text{from (3.234) or (3.235)}\\
T^s & \text{from (3.236) or (3.237)}
\end{array}\right\}\ 3 \quad \text{equations}
\end{array}
\tag{3.239}$$

The complexity of the governing equations is very high and a further reduction is useful and really possible in many applications. The reduction will be done in three levels in a top-down manner:

1. *First level reduction:* Multiphase variable-density flow, mass and heat transport in porous media based on the general Darcy-Brinkman-Forchheimer (DBF) flow equation.
2. *Second level reduction:* Single liquid phase variable-density flow, mass and heat transport in variably saturated porous media based on the Darcy flow equation [59].
3. *Third level reduction:* Variable-density Darcy-type flow, mass and heat transport in groundwater (fully saturated porous media), including vertically integrated formulations for aquifers.

3.9.2 Proper Reduction of Governing Equations for Multiphase Variable-Density Flow, Mass and Heat Transport in Porous Media: First Level Reduction

Introducing the volumetric flux density (*Darcy velocity*)[9] for the fluid phases $f = l, g$

$$\boldsymbol{q}_f = \varepsilon_f(\boldsymbol{v}^f - \boldsymbol{v}^s) = \varepsilon_f \boldsymbol{v}^{fs} \tag{3.240}$$

the general model Eqs. (3.225)–(3.238) can be significantly reduced if the following assumptions are made:

- Due to the generally slow motion of fluid flow in porous media the inertial effects appearing in the momentum conservation (3.231) or (3.232) in form of local acceleration $\partial(\varepsilon_f \rho^f \boldsymbol{v}^f)/\partial t$ and of convective acceleration $\nabla \cdot (\varepsilon_f \rho^f(\boldsymbol{v}^f \boldsymbol{v}^f))$ are negligible, cf. [389].
- Energy dissipation terms in the energy conservations equations (3.234)–(3.237) can be neglected: $\varepsilon_f p^f \nabla \cdot \boldsymbol{v}^f \approx 0$, $\frac{2}{3}\varepsilon_f \mu^f (\nabla \cdot \boldsymbol{v}^f)^2 \approx 0$, $2\varepsilon_f \mu^f \boldsymbol{d}^f{:}\boldsymbol{d}^f \approx 0$, $\varepsilon_s p^s \nabla \cdot \boldsymbol{v}^s \approx 0$, $\varepsilon_s(\boldsymbol{t}^s \cdot \boldsymbol{\epsilon}^s){:}\boldsymbol{\epsilon}^s \approx 0$.
- It is assumed that the phases of the porous medium are *locally* in a state of thermodynamic equilibrium. That means that the REV-averaged temperatures of all phases l, g, s are assumed to be equal at each point in the multiphase system:

  $$T^l = T^g = T^s = T \tag{3.241}$$

  where T represents the *system temperature*. As the consequence of (3.241) the energy conservation equations (3.234)–(3.237) can be summed up over all phases and only one energy equation for the multiphase systems finally results. Additionally, for the gas phase g the thermal capacity c^g and thermal hydrodynamic conductivity $\boldsymbol{\Lambda}^g$ can be neglected with respect to the solid and liquid phases. Another direct consequence of (3.241) is that the overall thermal conductivity $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_f + \boldsymbol{\Lambda}_s = \boldsymbol{\Lambda}_0 + \rho^f c^f \boldsymbol{D}_{f\,\mathrm{mech}}$ leads to a weighted arithmetic mean of the thermal conductivities of the fluid and solid phases in the form of $\boldsymbol{\Lambda}_0 = [\varepsilon s^f \Lambda^f + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta}$ as a natural result in which the thermal conductivities of the fluid and solid phases occur in *parallel*[10].

[9] Sometimes, the volumetric flux density is simply represented by the so-called *Dupuit-Forchheimer relationship* [389], which is a bulk flux in the form $\boldsymbol{v}_f = \varepsilon_f \boldsymbol{v}^f$. This quantity has been given various names by different authors (e.g., seepage or filtration velocity). We shall prefer the term *Darcy velocity* $\boldsymbol{q}_f$ emphasizing the correct relationship (3.240) for the flux.

[10] While a parallel behavior occurs in most of the natural porous media, there could be a porous-medium structure and orientation, where the heat conduction takes place in *series*. In this case, the heat flux can pass serially though the solid and the fluid, such that the overall thermal conductivity is a harmonic mean $\boldsymbol{\Lambda}_0^{-1} = \varepsilon s^f(\Lambda^f \boldsymbol{\delta})^{-1} + (1-\varepsilon)(\Lambda^s \boldsymbol{\delta})^{-1}$. The arithmetic mean and harmonic mean represent upper and lower bounds, respectively, for the overall thermal conductivity $\boldsymbol{\Lambda}_0$.

- The solid phase s is assumed deformable, but solid grains are incompressible. Inserting (3.240) into (3.225) and using the definition (3.219) the mass conservation equation for the fluid phase f reads

$$\varepsilon s^f \frac{\partial \rho^f}{\partial t} + \rho^f s^f \frac{\partial \varepsilon}{\partial t} + \rho^f \varepsilon \frac{\partial s^f}{\partial t} + \nabla \cdot (\rho^f \boldsymbol{q}_f) + \nabla \cdot (\varepsilon s^f \rho^f \boldsymbol{v}^s) = \rho^f Q_f \tag{3.242}$$

Assuming slowly deformable media and slightly compressible fluids the following approximation holds [37]

$$\nabla \cdot (\varepsilon s^f \rho^f \boldsymbol{v}^s) \approx \varepsilon s^f \rho^f (\nabla \cdot \boldsymbol{v}^s) \tag{3.243}$$

The expression $\nabla \cdot \boldsymbol{v}^s$ is obtained from the solid mass balance (3.226)

$$\frac{\partial [(1-\varepsilon)\rho^s}{\partial t} + \nabla \cdot [(1-\varepsilon)\rho^s \boldsymbol{v}^s] = 0 \tag{3.244}$$

where $Q_s = 0$ is assumed. For incompressible solid grains, (3.244) becomes

$$\nabla \cdot \boldsymbol{v}^s \approx \left(\frac{1}{1-\varepsilon} \right) \frac{\partial \varepsilon}{\partial t} \tag{3.245}$$

In changing the porosity ε of the porous-medium compression work of the skeleton is taken into account. Let us consider the porosity as a function of fluid pressure and let mass fraction and temperature effects be disregarded, we have the differential

$$d\varepsilon = \frac{\partial \varepsilon}{\partial p^f} dp^f = \underbrace{\left(\frac{1}{1-\varepsilon} \frac{\partial \varepsilon}{\partial p^f} \right)}_{\upsilon} (1-\varepsilon) dp^f = \upsilon (1-\varepsilon) dp^f \tag{3.246}$$

where υ represents the coefficient of *skeleton compressibility*. It takes into account a vertical deformation of the porous medium. The relations (3.245) and (3.246) decouple the fluid equations from the solid equations and there is no need anymore to solve explicitly the momentum conservation equation (3.233) and mass conservation equation (3.226) for the solid s. This approach is a common practice in subsurface modeling, where the movement of the solid phase is modeled only implicitly. Rather than trying to obtain detailed information about movement of the solid phase, only its compression is considered. This assumption can be inappropriate for problems of land subsidence, slope or embankment stability in a geotechnical context [344] or for large deformations in absorbent swelling industrial porous material [144, 147, 375].

Other, more empirical arrangements for $\boldsymbol{\Lambda}_0$ can be made up for certain porous media as discussed in [305].

- For the species mass, momentum and energy conservation equations the velocity term $\varepsilon_f \boldsymbol{v}^f$ can be replaced by the volumetric flux density $\boldsymbol{q}_f$ (3.240), assuming that the terms associated with the solid movement $\varepsilon_f \boldsymbol{v}^s$ are negligible.
- Let us consider a species k, which occurs both in the fluid phase f and the solid phase s. Let us assume that k is sorbed at the solid phase, which can be expressed by the *sorption isotherm* (for more details see Chap. 5):

$$\rho^s \omega_k^s = \rho^f \varphi_k \, \omega_k^f \tag{3.247}$$

where $\varphi_k = \varphi_k(\omega_k^f)$ is the dimensionless *adsorption function*, which can be dependent on ω_k^f. In such an adsorption process a fluid phase f can occupy only part of the void space ε and therefore only part of the total area of the solid can be exposed to adsorption [39]. Sometimes, it is assumed [38, 422] that the wetting phase completely coats the solid such that no other (nonwetting) fluid phase is in contact with the solid. To take into account this solid-fluid contact phenomenon for the adsorption of chemical species k occurring both in the wetting fluid phase f and in the absorbing solid phase s, we subdivide the solid volume fraction ε_s into chemically active and inactive parts of solid mass $\varepsilon_s = \varepsilon_{s^{\text{active}}} + \varepsilon_{s^{\text{inactive}}}$ (cf. Sect. 5.2.2). Accordingly, the mass balances (3.227) and (3.229) (assuming $\nabla \cdot (\varepsilon_s \rho^s \boldsymbol{v}^s \omega_k^s) \approx 0$) of species k in both phases f and s can be written as

$$\begin{aligned} \frac{\partial}{\partial t}(\varepsilon s^f \rho^f \omega_k^f) + \nabla \cdot (\rho^f \boldsymbol{q}_f \omega_k^f) + \nabla \cdot \boldsymbol{j}_k^f &= \varepsilon s^f (r_k^f - \vartheta_k \omega_k^f + \tilde{R}_k^f) \\ \frac{\partial}{\partial t}(\varepsilon_{s^{\text{active}}} \rho^s \omega_k^s) &= \varepsilon_{s^{\text{active}}} (r_k^s - \vartheta_k \omega_k^s + \tilde{R}_k^s) \end{aligned} \tag{3.248}$$

where in the heterogeneous reaction rates R_k^f and R_k^s the linear reaction parts of *decay* are split off according to $R_k^f = -\vartheta_k \rho^f \omega_k^f + \tilde{R}_k^f$, $R_k^s = -\vartheta_k \rho^s \omega_k^s + \tilde{R}_k^s$, introducing a joint linear decay rate constant ϑ_k of species k (see Sect. 5.4.2). The ratio of the area of the adsorbing solid-fluid interface to the total area of the solid, which can be assumed equal to the ratio of active (adsorbing) solid mass to the total mass of solid, and accordingly assumed equal to the ratio of the active (adsorbing) solid mass fraction $\varepsilon_{s^{\text{active}}}$ to the total solid mass fraction ε_s, is apparently a function of the saturation s^f of the wetting fluid phase: $\varepsilon_{s^{\text{active}}}/\varepsilon_s = f(s^f)$, where $s^f \leq f(s^f) \leq 1$ is a surface contact ratio function, which has to be specified. In many applications a suited approximation is $f(s^f) \approx s^f$ and we use

$$\varepsilon_{s^{\text{active}}} = f(s^f)\varepsilon_s \approx s^f(1 - \varepsilon) \tag{3.249}$$

Inserting (3.247) and (3.249) into (3.248), we can add up the mass conservation equations (3.248) to obtain:

$$\begin{aligned} \frac{\partial}{\partial t}(\varepsilon s^f \rho^f \Re_k \omega_k^f) + \nabla \cdot (\rho^f \boldsymbol{q}_f \omega_k^f) + \nabla \cdot \boldsymbol{j}_k^f + \varepsilon s^f \rho^f \vartheta_k \Re_k \omega_k^f \\ = \underbrace{s^f \Big[\varepsilon(r_k^f + \tilde{R}_k^f) + (1 - \varepsilon)(r_k^s + \tilde{R}_k^s)\Big]}_{\tilde{R}_k} \end{aligned} \tag{3.250}$$

written in the divergence form and

$$\begin{aligned}\varepsilon s^f \rho^f \acute{\Re}_k \frac{\partial \omega_k^f}{\partial t} + \rho^f \boldsymbol{q}_f \cdot \nabla \omega_k^f + \nabla \cdot \boldsymbol{j}_k^f + \varepsilon s^f \rho^f \vartheta_k \Re_k \omega_k^f \\ = \tilde{R}_k - \rho^f \omega_k^f Q_f \end{aligned} \qquad (3.251)$$

written in the convective form, where $\Re_k$ is the *retardation factor*

$$\Re_k = 1 + \Big(\frac{1-\varepsilon}{\varepsilon}\Big)\varphi_k \qquad (3.252)$$

and $\acute{\Re}_k$ is the derivative term of retardation

$$\acute{\Re}_k = 1 + \Big(\frac{1-\varepsilon}{\varepsilon \rho^f}\Big)\frac{\partial(\rho^f \varphi_k \omega_k^f)}{\partial \omega_k^f} \qquad (3.253)$$

- As stated in Sect. 3.7.3, the summation of the species mass balance equations over all $N = \sum_\alpha N^\alpha$ species must give the total mass balance of the phase(s), so that only $N^* = \sum_\alpha (N^\alpha - 1)$ of the species mass fractions are independent, because if $N^\alpha - 1$ are known, the N^αth may be computed directly from $\omega_{N^\alpha}^\alpha = 1 - \sum_{k=1}^{N^\alpha - 1} \omega_k^\alpha$. Accordingly, only N^* species mass transport equations are needed to be solved, where N^* denotes the *essential number of species*.

Taking into account the above assumptions we find the governing equations of the *first level reduction* as summarized in Table 3.5. In the momentum equations three terms are emphasized which are of specific concern.

First, the Brinkman term $\frac{\mu^f}{\varepsilon_f}\nabla^2 \boldsymbol{q}_f$ results from the viscous shear stresses of the fluid. Brinkman (see [389, 534] for references) has firstly described this term in the context of porous media, but, had set the term to $\mu^f \nabla^2 \boldsymbol{q}_f$. However, the correct factor must be $\frac{\mu^f}{\varepsilon_f}$ instead of μ^f resulting directly from by the present volume averaging procedure, cf. [394]. It was pointed out by Tam [505] that whenever the length scale of the investigated problem is much greater than $(\|\boldsymbol{k}\|/\varepsilon)^{1/2}$ the Brinkman term becomes negligible in comparison to the Darcy term. Only for thin boundary layers with a thickness lower than $(\|\boldsymbol{k}\|/\varepsilon)^{1/2}$ the Brinkmann term could have effects for practical applications.

Second, the Darcy term $\mu^f (k_r^f \boldsymbol{k})^{-1} \cdot \boldsymbol{q}_f$ represents a linear relationship to $\boldsymbol{q}_f$ due to viscous drag by friction at the solid-fluid interfaces of the porous medium. This holds when $\boldsymbol{q}_f$ is sufficiently small, which is valid for most of the porous-medium applications. The characteristic measure is provided by the *pore Reynolds number* Re_p of the flow defined by

$$\mathrm{Re}_p = \frac{\|\boldsymbol{q}_f\| \rho^f d}{\mu^f} \qquad (3.254)$$

Table 3.5 Summarized balance laws and constitutive relations (CR) of multiphase variable-density DBF-type flow, mass and heat transport in porous media as *first level* model reduction. It forms a system of $3 + N^* + 2D$ equations[a] to solve the (3) variables p^l, p^g, T, the (N^*) variables[b] ω_k^f of species k (or ω_m^s of species m) in the fluid phases f and in the solid phase s, respectively, and the (2D) variables $\boldsymbol{q}_l$ and $\boldsymbol{q}_g$. Alternative convective forms are given in angle brackets.

Type	Equations
$\mathcal{M}^f$	$\varepsilon s^f \frac{\partial \rho^f}{\partial t} + \rho^f \left(s^f \upsilon \frac{\partial p^f}{\partial t} + \varepsilon \frac{\partial s^f}{\partial t} \right) + \nabla \cdot (\rho^f \boldsymbol{q}_f) = \rho^f Q_f$
$\mathcal{M}_k^f + \mathcal{M}_k^s$	$\frac{\partial}{\partial t}(\varepsilon s^f \rho^f \mathfrak{R}_k \omega_k^f) + \nabla \cdot (\rho^f \boldsymbol{q}_f \omega_k^f) + \nabla \cdot \boldsymbol{j}_{fk} + \varepsilon s^f \rho^f \vartheta_k \mathfrak{R}_k \omega_k^f = \tilde{R}_k$ $\left\langle \varepsilon s^f \rho^f \acute{\mathfrak{R}}_k \frac{\partial \omega_k^f}{\partial t} + \rho^f \boldsymbol{q}_f \cdot \nabla \omega_k^f + \nabla \cdot \boldsymbol{j}_{fk} + \varepsilon s^f \rho^f \vartheta_k \mathfrak{R}_k \omega_k^f = \tilde{R}_k - \rho^f \omega_k^f Q_f \right\rangle$
$\mathcal{M}_m^s$	$\frac{\partial}{\partial t}(\varepsilon_s \rho^s \omega_m^s) = \varepsilon_s (r_m^s + R_m^s) \qquad m \neq k$ $\left\langle \varepsilon_s \rho^s \frac{\partial \omega_m^s}{\partial t} = \varepsilon_s (r_m^s + R_m^s) \right\rangle$
$\mathcal{V}^f$	$\mathbf{0} = -\nabla p^f + \rho^f \boldsymbol{g} + \underbrace{\frac{\mu^f}{\varepsilon s^f} \nabla^2 \boldsymbol{q}_f}_{\text{Brinkman}} - \underbrace{\mu^f (k_r^f \boldsymbol{k})^{-1} \cdot \boldsymbol{q}_f}_{\text{Darcy}} - \underbrace{\rho^f (k_r^f \boldsymbol{k})^{-1/2} c_F \lVert \boldsymbol{q}_f \rVert \cdot \boldsymbol{q}_f}_{\text{Forchheimer}}$
$\mathcal{E}^l + \mathcal{E}^s$	$\frac{\partial}{\partial t}\Big[\big(\varepsilon s^l \rho^l c^l + (1-\varepsilon)\rho^s c^s\big)(T - T_0)\Big] + \nabla \cdot (\rho^l c^l \boldsymbol{q}_l (T - T_0)) - \nabla \cdot (\boldsymbol{\Lambda} \cdot \nabla T) = H_e$ $\left\langle \big(\varepsilon s^l \rho^l c^l + (1-\varepsilon)\rho^s c^s\big) \frac{\partial T}{\partial t} + \rho^l c^l \boldsymbol{q}_l \cdot \nabla T - \nabla \cdot (\boldsymbol{\Lambda} \cdot \nabla T) = H_e - \rho^l c^l (T - T_0) Q_l \right\rangle$
CR	$\rho^f = \rho_0^f \big[1 + \gamma^f (p^f - p_0^f) + \sum_{k=1}^{N^f-1} \alpha_k^f (\omega_k^f - \omega_{k0}^f)$ $\quad -\beta^f (T)(T - T_0)\big]$ $p_c = p_c(s^l) = p^g - p^l$ $k_r^f = k_r^f (s^f)$ $\boldsymbol{j}_{fk}(\Im_H \lVert \boldsymbol{j}_{fk} \rVert + 1) = -\rho^f \boldsymbol{D}_{fk} \cdot \nabla \omega_k^f$ $\boldsymbol{D}_{fk} = \varepsilon s^f D_k^f \boldsymbol{\delta} + \boldsymbol{D}_{f\text{mech}}$ $\boldsymbol{D}_{f\text{mech}} = \beta_T^f \lVert \boldsymbol{q}_f \rVert \boldsymbol{\delta} + (\beta_L^f - \beta_T^f) \frac{\boldsymbol{q}_f \otimes \boldsymbol{q}_f}{\lVert \boldsymbol{q}_f \rVert}$ $\mathfrak{R}_k = 1 + \left(\frac{1-\varepsilon}{\varepsilon}\right) \varphi_k$ $\acute{\mathfrak{R}}_k = 1 + \left(\frac{1-\varepsilon}{\varepsilon \rho^f}\right) \frac{\partial(\rho^f \varphi_k \omega_k^f)}{\partial \omega_k^f}$ $\boldsymbol{\Lambda} = \big[\varepsilon s^l \Lambda^l + (1-\varepsilon) \Lambda^s\big] \boldsymbol{\delta} + \rho^l c^l \boldsymbol{D}_{l\text{mech}}$ $\mu^f = \mu^f (\omega_k^f, T)$ $H_e = \rho^l H_l + \rho^s H_s$

[a] $f = l, g$.

[b] Species k can occur both in the fluid phase f and the solid phase s, however, species $m \neq k$ only occurs in the solid phase s.

where d is the characteristic length dimension representing the elementary channels of the porous medium. It could be used as a mean grain diameter or estimated via $(\lVert \boldsymbol{k} \rVert / \varepsilon)^{1/2}$. The linear drag of the Darcy flow regime is valid as long as Re_p does not exceed some values between 1 and 10

Table 3.6 Types of momentum equations $\mathcal{V}^f$ for the fluid phases f

Type	Formulations[a]
Darcy-Brinkman-Forchheimer	$\boldsymbol{q}_f\Big(c_F\frac{\rho^f}{\mu^f}(k_r^f\,\|\boldsymbol{k}\|)^{1/2}\|\boldsymbol{q}_f\|+1\Big)=-\frac{k_r^f\,\boldsymbol{k}}{\mu^f}\cdot(\nabla p^f-\rho^f\boldsymbol{g})+\frac{k_r^f\,\boldsymbol{k}}{\varepsilon s^f}\cdot\nabla^2\boldsymbol{q}_f$
Darcy-Forchheimer	$\boldsymbol{q}_f\Big(c_F\frac{\rho^f}{\mu^f}(k_r^f\,\|\boldsymbol{k}\|)^{1/2}\|\boldsymbol{q}_f\|+1\Big)=-\frac{k_r^f\,\boldsymbol{k}}{\mu^f}\cdot(\nabla p^f-\rho^f\boldsymbol{g})$
Darcy	$\boldsymbol{q}_f=-\frac{k_r^f\,\boldsymbol{k}}{\mu^f}\cdot(\nabla p^f-\rho^f\boldsymbol{g})$

[a] The square root of $(k_r^f\,\boldsymbol{k})^{1/2}$ is usually approximated by $(k_r^f\,\|\boldsymbol{k}\|)^{1/2}\boldsymbol{\delta}$

$$\mathrm{Re}_p < 1\ldots 10 \tag{3.255}$$

However, as $\boldsymbol{q}_f$ (particularly Re_p) increases the viscous drag becomes nonlinear.

Third, a quadratic drag term is provided by the Forchheimer term $\rho^f\,(k_r^f\,\boldsymbol{k})^{-1/2}$ $c_F\|\boldsymbol{q}_f\|\cdot\boldsymbol{q}_f$, which takes into account that the transition from linear to nonlinear drag is smooth and does not mean that there is a sudden transition from laminar to turbulent flow. Even in laminar flow regimes the linearity is broken due to the fact that the form drag due to solid obstacles is now comparable with the surface drag due to friction. An upper limit of Re_p at about 100 is suggested [33] for the nonlinear laminar flow regime. For higher Re_p−numbers a turbulent flow regime occurs, which requires the extension to turbulent transport mechanisms based on the full momentum equations [118]. In the Forchheimer term we introduced the more common dimensionless form-drag constant c_F (3.155), which is often approximated by 0.55. In dependence on the meaning of the different terms in the momentum equation $\mathcal{V}^f$ of Table 3.5 we differ between (1) the Darcy-Brinkman-Forchheimer (DBF) equation, (2) the Darcy-Forchheimer (DF) equation and (3) the Darcy equation as listed in Table 3.6.

3.10 Final Model Equations for Flow, Mass and Heat Transport

3.10.1 Single Liquid Phase Variable-Density Flow, Mass and Heat Transport in Variably Saturated Porous Media: Second Level Reduction

For the practical modeling of variable-density flow, mass and heat transport in porous media the above balance laws with their constitutive relations of the first level reduction as listed in Tables 3.5 and 3.6 are further simplified. The following assumptions are made as the *second level* reduction of the governing model equations:

- *Fluid phase assumption:* In the void space two fluid phases f coexist: a liquid phase l (e.g., water) and a gas phase g (e.g., air). In many applications, however, the gas phase g can be assumed *stagnant*, i.e.,

$$\boldsymbol{q}_g = \varepsilon_g(\boldsymbol{v}^g - \boldsymbol{v}^s) \equiv \boldsymbol{0} \tag{3.256}$$

  Accordingly, there is no need to consider anymore the momentum balance for the gas phase. As a consequence of (3.256) a *hydrostatic* gas pressure condition with $\nabla p^g = \rho^g \boldsymbol{g}$ results from the momentum balance equations for the gas phase (Table 3.6) and p^g could be solved explicitly as a simple function of gas density ρ^g and location $\boldsymbol{x}$. This assumption reduces the problem to a single-phase flow, where the only dynamic fluid phase is the liquid phase l, however, under variable liquid saturation s^l of the void space ε, for which a further assumption is required to decouple finally the liquid phase from the gas phase.
- *Capillary pressure assumption:* The liquid saturation s^l is determined from the capillary pressure (3.223): $p_c(s^l) = p^g - p^l$. Taking into account that the density of liquid is much higher than of gas (e.g., note that the relation between water and air is $\rho^l/\rho^g \approx 800$), with the hydrostatic gas phase condition (3.256) we find that $\nabla p^l \gg \rho^g \boldsymbol{g}$ and conclude that gravitational effects on the gas pressure p^g are negligible in comparison to the liquid pressure p^l. This allows us to assume a constant gas pressure $p^g \approx$ const. Practically, we refer to a constant atmospheric pressure and set $p^g = 0$. It simplifies the capillary pressure relation according to

$$p_c(s^l) = -p^l \tag{3.257}$$

  With this assumption the liquid phase is actually decoupled from the gas phase and the flow and transport process of the liquid phase may be modeled without need to explicitly model the gas phase. It represents the key assumption of flow and transport modeling in variably saturated porous media. Based on the relationships as described in Appendix D it is now easy to relate directly the liquid pressure to the liquid saturation $p^l = p^l(s^l)$ and, inversely, to express the liquid saturation as a function of the liquid pressure $s^l = s^l(p^l)$.
- *Momentum equation assumption:* It can be usually assumed that the liquid phase moves slowly in the porous medium and the condition (3.255) is satisfied. Accordingly, the momentum balance for the fluid phase l can be described by the Darcy equation (Table 3.6):

$$\boldsymbol{q}_l = -\frac{k_r^l \boldsymbol{k}}{\mu^l} \cdot (\nabla p^l - \rho^l \boldsymbol{g}) \tag{3.258}$$

3.10.2 Choice of Suited Variables

In groundwater hydraulics and subsurface hydrology it is common to measure pressures at a point P above a reference datum in an equivalence to a head of liquid

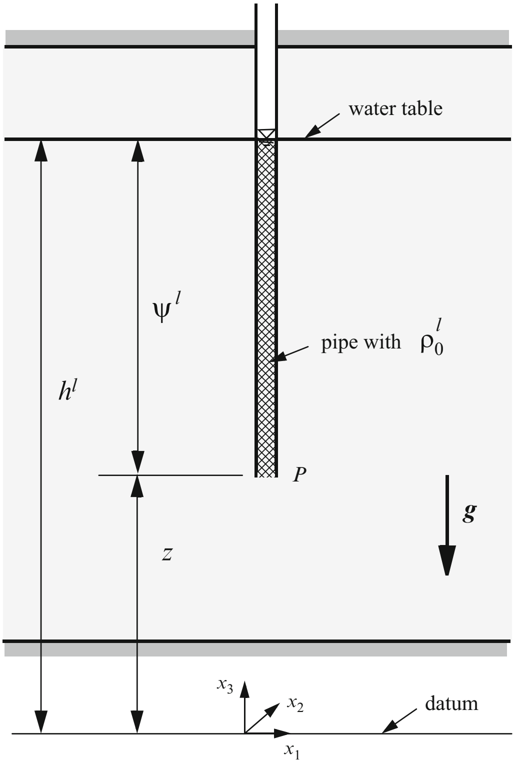


Fig. 3.12 Pressure head $\psi^l = p^l/(\rho_0^l g)$ and hydraulic head $h^l = \psi^l + z$ measured in a piezometric pipe

(e.g., water) with given density in a vertical column (e.g., pipe, well) as shown in Fig. 3.12. We define the *pressure head* ψ^l of the liquid l

$$\psi^l = \frac{p^l}{\rho_0^l g} \tag{3.259}$$

and the *hydraulic head* (piezometric head) h^l of the liquid l

$$h^l = \frac{p^l}{\rho_0^l g} + x_j = \psi^l + x_j \tag{3.260}$$

which are related to the constant reference liquid density ρ_0^l, where the subscript $j = 1, 2$ or 3 indicates the direction of gravity aligned to a major coordinate direction of $\boldsymbol{x}$. Typically, in a vertical direction it is $x_j = x_3 = z$ and $\boldsymbol{g}^T = (0\ 0\ g)$, where $g = \|\boldsymbol{g}\|$ is the gravitational acceleration. Introducing the gravitational unit vector

$$\boldsymbol{e} = -\frac{\boldsymbol{g}}{g} \quad (= \nabla x_j) \tag{3.261}$$

we can express the Darcy equation (3.258) by the variables of hydraulic head h^l and pressure head ψ^l, respectively[11]

$$\begin{aligned} \boldsymbol{q}_l &= -k_r^l \boldsymbol{K}^l f_\mu^l \cdot \left(\nabla h^l + \chi^l \boldsymbol{e}\right) \\ \boldsymbol{q}_l &= -k_r^l \boldsymbol{K}^l f_\mu^l \cdot \left[\nabla \psi^l + (1+\chi^l)\boldsymbol{e}\right] \end{aligned} \tag{3.262}$$

where

$$\boldsymbol{K}^l = \frac{\boldsymbol{k}\rho_0^l g}{\mu_0^l} \tag{3.263}$$

defines the *hydraulic conductivity*,

$$f_\mu^l = \frac{\mu_0^l}{\mu^l} \tag{3.264}$$

is the *viscosity relation function* (3.218) and

$$\chi^l = \frac{\rho^l - \rho_0^l}{\rho_0^l} \tag{3.265}$$

is the dimensionless *buoyancy coefficient* of the liquid phase l.

It is important to note that the hydraulic conductivity $\boldsymbol{K}^l$ incorporates both porous-medium and liquid properties, however, the liquid parameters ρ_0^l and μ_0^l in (3.263) represent constant reference values and accordingly $\boldsymbol{K}^l$ remains *de facto* a parameter of the porous medium scaled with constant liquid parameters ρ_0^l and μ_0^l and the gravitational constant g. Through the Darcy equation (3.262) formulated with the hydraulic conductivity $\boldsymbol{K}^l$, the actual pressure, species concentration and temperature effects on the liquid density ρ^l and liquid viscosity μ^l are implied by the buoyancy coefficient χ^l (3.265) with (3.199) and the viscosity relation function f_μ^l (3.218), respectively. Clearly, the $h/\psi-$formulations (3.262) are fully physically equivalent to the basic $p-$formulation (3.258) of the Darcy equation.

[11] From (3.260) it is $p^l = \rho_0^l g(h^l - x_j)$ and with $\boldsymbol{e} = \nabla x_j$ we find $\nabla p^l = \rho_0^l g(\nabla h^l - \boldsymbol{e})$. Now expanding

$$\frac{\boldsymbol{k}}{\mu^l} = \underbrace{\frac{\boldsymbol{k}\rho_0^l g}{\mu_0^l}}_{\boldsymbol{K}^l} \underbrace{\frac{\mu_0^l}{\mu^l}}_{f_\mu^l} \frac{1}{\rho_0^l g} = \boldsymbol{K}^l f_\mu^l \frac{1}{\rho_0^l g}$$

and inserting into (3.258) with (3.261), we obtain

$$\boldsymbol{q}_l = -k_r^l \boldsymbol{K}^l f_\mu^l \cdot \left(\nabla h^l + \tfrac{\rho^l - \rho_0^l}{\rho_0^l}\boldsymbol{e}\right)$$

In the above species mass balance equations of Table 3.5 the dimensionless mass fraction ω_k^α appears as the natural variable of mass conservation. In practice, however, the mass concentration C_k^α (2.117) is often preferred, which is related to the mass fraction ω_k^α according to (2.123)

$$C_k^\alpha = \rho^\alpha \, \omega_k^\alpha \tag{3.266}$$

The replacement of mass fraction by mass concentration in the species mass conservation equations requires for some specific terms a further consideration, which is part of the next subject.

3.10.3 Oberbeck-Boussinesq Approximation and Extension

The system of balance equations listed in Table 3.5 is coupled by the nonlinearity in the fluid density ρ^l. Its analysis can be substantially simplified by the so-called *Oberbeck-Boussinesq (OB) approximation*, sometimes termed only as Boussinesq approximation. As pointed out in [255,389] the term OB approximation seems more appropriate because Oberbeck [393] addressed this problem before Boussinesq [49].

The OB approximation consists in neglecting all density dependencies in the balance terms, except for the crucial buoyancy term $\rho^l \boldsymbol{g}$ (or $\chi^l \boldsymbol{e}$) which is retained in the momentum equation of Table 3.6 (or (3.262)). For the buoyancy term the fluid density dependency (3.199) is incorporated as a function of mass fraction ω_k^l and temperature T, however, no pressure dependency is considered here. Pressure dependency remains a subject of the derivative term $\partial\rho^l/\partial t$ appearing in the LHS of the liquid mass balance equation as further discussed below. Referring to saturated and nondeformable porous media and considering liquid incompressibility as well as no sources/sinks, the liquid mass conservation $\mathcal{M}^l$ of Table 3.5 reduces then to the simple expression $\nabla \cdot \boldsymbol{q}_l = 0$ and the velocity becomes solenoidal, cf. Sect. 2.1.10. This incompressibility assumption is common in most analytical and stability analyses of convection phenomena.

The OB approximation is valid if density changes $\Delta\rho^l$ remain small in comparison to the reference density ρ_0^l. Criteria for the validity of the OB approximation for liquids and gases were given by Gray and Giorgini [204]. Obviously, the OB approximation becomes invalid for large density variations, e.g., at high-concentration brines and/or high temperature gradients. However, it is often not clear what consequences practically result if the full dependencies are incorporated (so-called *non-Boussinesq effects*). Usually, extensions to non-Boussinesq formulations can be introduced by 'correction' terms written for the liquid mass conservation equation $\mathcal{M}^l$ of Table 3.5 in the following form

$$\frac{\varepsilon s^l}{\rho^l}\frac{\partial \rho^l}{\partial t}\bigg|_{T,\omega_k^l} + s^l \upsilon \frac{\partial p^l}{\partial t} + \varepsilon \frac{\partial s^l}{\partial t} + \nabla \cdot \boldsymbol{q}_l = Q_l + Q_{l_{\mathrm{EOB}}} \tag{3.267}$$

with the extended Boussinesq approximation term

$$Q_{l_{\mathrm{EOB}}} = -\frac{1}{\rho^l}\Big(\boldsymbol{q}_l \cdot \nabla\rho^l + \varepsilon s^l \frac{\partial\rho^l}{\partial t}\Big|_{p^l}\Big) \tag{3.268}$$

where $|_{T,\omega_k^l}$ and $|_{p^l}$ indicate that T, ω_k^l and p^l, respectively, are held constant. Inserting the EOS for the liquid density (3.199) into (3.268) we can approximate

$$Q_{l_{\mathrm{EOB}}} = -\boldsymbol{q}_l \cdot \Big(\gamma^l \nabla p^l + \sum_k \alpha_k^l \nabla\omega_k^l - \beta^{l*}\nabla T\Big) - \varepsilon s^l \Big(\sum_k \alpha_k^l \frac{\partial\omega_k^l}{\partial t} - \beta^{l*}\frac{\partial T}{\partial t}\Big) \tag{3.269}$$

introducing a generalized thermal expansion coefficient β^{l*}

$$\beta^{l*} = \begin{cases} \beta^l & \text{for constant expansion} \\ \dfrac{\beta^l(T) + \frac{\partial\beta^l(T)}{\partial T}(T - T_0)}{1 + \sum_k \alpha_k^l(\omega_k^l - \omega_{k0}^l) - \beta^l(T)(T - T_0)} & \text{for variable expansion} \end{cases} \tag{3.270}$$

where β^l is a given constant, while $\beta^l(T)$ and $\partial\beta^l(T)/\partial T$ correspond to (C.8) and (C.10), respectively, derived in Appendix D.

Kolditz et al. [318] compared OB solutions and some extended forms exemplified for the Elder cellular convection problem (cf. Sect. 11.11.4). For this case, OB solutions were rather close to non-Boussinesq model results. Only slight differences in pressure and concentration distributions in some parts of the model domain were observed. Evans and Raffensperger [160] studied the limitation of the OB approximation for a problem which is similar to the Elder problem. They found differences in the concentration distributions up to 9 % comparing the results of the different formulations. Gartling and Hickox [186] studied adjustments for the variation of fluid properties in the heat transport equation, while assuming the constraint of incompressibility, $\nabla \cdot \boldsymbol{q}_l = 0$. They found that the OB approximation and their extended solutions can be sufficiently 'close' for integrated quantities over large temperature ranges. However, differences can occur for local quantities. The accurate prediction of the flow field has been shown to be of major concern, and they concluded that the 'goodness' of the OB solutions depends on what quantities are of interest in the problem solution.

Furthermore, it is to be noted that under large compression effects when γ^l becomes significant the OB solution can considerably violate mass conservativity and the extended OB approximation is to be preferred. However, for a realistically small liquid compressibility γ^l, the term $-\boldsymbol{q}_l \cdot \gamma^l \nabla p^l$ in (3.269) is usually negligible.

If we replace the mass fraction ω_k^l by the mass concentration C_k^l for species k in the governing equations of Table 3.5, we have to neglect density variations in the *convective* form of the species mass transport equation (3.228) and in the species mass flux vector $\boldsymbol{j}_{lk}$, (3.187) or (3.183). These assumptions are acceptable within the OB approximation and its extension. This allows to approximate the derivative terms in the convective form as

$$\rho^l \frac{\partial \omega_k^l}{\partial t} \approx \frac{\partial C_k^l}{\partial t}, \quad \rho^l \nabla \omega_k^l \approx \nabla C_k^l \tag{3.271}$$

assuming $(C_k^l/\rho^l)\partial\rho^l/\partial t \approx 0$, $(C_k^l/\rho^l)\nabla\rho^l \approx \mathbf{0}$, and to write the species mass flux vector $\boldsymbol{j}_{lk}$ in the form

$$\boldsymbol{j}_{lk}(\Im_H \|\boldsymbol{j}_{lk}\| + 1) = -\boldsymbol{D}_{lk} \cdot \nabla C_k^l \tag{3.272}$$

assuming $(C_k^l/\rho^l)\boldsymbol{D}_{lk} \cdot \nabla\rho^l \approx \mathbf{0}$. Note that an evident advantage results in the *divergence* form (3.227) if ω_k^l is replaced by C_k^l because assumption (3.271) is not necessary anymore.

Similarly, within the OB approximation and its extension density variations in the terms of the governing heat transport equations, both for the divergence and convective form, are neglected too.

3.10.4 Reformation of Terms

With the replacement of the pressure variable p^l (3.108) by the hydraulic head h^l (3.260) (or pressure head ψ^l (3.259)) and the species mass fraction ω_k^α (2.123) by the mass concentration C_k^α (2.117) we have to adjust specific terms in the governing equations of Table 3.5. First, the differential of the liquid density ρ^l (3.197) is modified:

$$\begin{aligned} d\rho^l &= \gamma^l \rho^l dp^l + \textstyle\sum_k \alpha_k^l \rho^l d\omega_k^l - \beta^l \rho^l dT \\ &= \gamma^l \frac{\partial p^l}{\partial h^l} \rho^l dh^l + \textstyle\sum_k \alpha_k^l \frac{\partial \omega_k^l}{\partial C_k^l} \rho^l dC_k^l - \beta^l \rho^l dT \\ &= \gamma^l \rho_0^l g \rho^l dh^l + \textstyle\sum_k \frac{\alpha_k^l}{C_{ks}^l - C_{k0}^l} \rho^l dC_k^l - \beta^l \rho^l dT \end{aligned} \tag{3.273}$$

to obtain

$$\rho^l = \rho_0^l \Big[1 + \gamma^l \rho_0^l g\,(h^l - h_0^l) + \sum_{k=1}^{N^l-1} \frac{\alpha_k^l}{C_{ks}^l - C_{k0}^l}(C_k^l - C_{k0}^l) - \beta^l(T)(T - T_0)\Big] \tag{3.274}$$

where h_0^l and C_{k0}^l are reference values of the hydraulic head and mass concentration of species k, respectively, and C_{ks}^l represents a given maximum mass concentration of species k, which may be used to estimate the specific solutal expansion coefficient by a linear relation

$$\alpha_k^l = \frac{\rho^l(C_{ks}^l) - \rho_0^l}{\rho_0^l} \tag{3.275}$$

sometimes called *density ratio*. A reasonable guess of the liquid density ρ^l at maximum concentration C_{ks}^l is

$$\rho^l(C_{ks}^l) \approx \rho_0^l + a\, C_{ks}^l \tag{3.276}$$

where Baxter and Wallace [32] proposed for the factor $a = 0.7$ and INTRAVAL project studies [395] used $a = 0.6923$. It gives an estimation of the specific solutal expansion coefficient according to

$$\alpha_k^l \approx \frac{a\, C_{ks}^l}{\rho_0^l} \tag{3.277}$$

The buoyancy coefficient (3.265) appearing in the Darcy equation (3.262) takes now with (3.274) the form:

$$\chi^l = \sum_{k=1}^{N^l-1} \beta_{c_k}^l (C_k^l - C_{k0}^l) - \beta^l(T)(T - T_0) \tag{3.278}$$

where we introduce with

$$\beta_{c_k}^l = \frac{\alpha_k^l}{C_{ks}^l - C_{k0}^l} \tag{3.279}$$

the *solutal expansion coefficient* of species k. The mass conservation equation of the liquid phase in the formulation of (3.267) can now be written as

$$\frac{\varepsilon s^l}{\rho^l}\frac{\partial \rho^l}{\partial h^l}\frac{\partial h^l}{\partial t} + s^l \upsilon \frac{\partial p^l}{\partial h^l}\frac{\partial h^l}{\partial t} + \varepsilon \frac{\partial s^l}{\partial t} + \nabla \cdot \boldsymbol{q}_l = Q_l + Q_{l_{\text{EOB}}} \tag{3.280}$$

to obtain by using (3.274) and (3.260)

$$s^l \underbrace{\rho_0^l g\big(\varepsilon\gamma^l + \upsilon\big)}_{S_o^l} \frac{\partial h^l}{\partial t} + \varepsilon \frac{\partial s^l}{\partial t} + \nabla \cdot \boldsymbol{q}_l = Q_l + Q_{l_{\text{EOB}}} \tag{3.281}$$

where $S_o^l = \rho_0^l g(\varepsilon\gamma^l + \upsilon)$ is the *specific storage coefficient*, sometimes called *specific storativity* [38], due to liquid and medium compressibility, and the correction sink term for the EOB approximation (3.268) gives now

$$Q_{l_{\mathrm{EOB}}} = -\boldsymbol{q}_l \cdot \Big(\gamma^l \rho_0^l g \nabla h^l + \sum_{k=1}^{N^l-1} \beta_{c_k}^l \nabla C_k^l - \beta^{l*} \nabla T\Big) - \\ \varepsilon s^l \Big(\sum_{k=1}^{N^l-1} \beta_{c_k}^l \frac{\partial C_k^l}{\partial t} - \beta^{l*} \frac{\partial T}{\partial t}\Big) \qquad (3.282)$$

To enforce conservativity for any magnitude of the specific storativity $S_o^l \geq 0$, in the EOB approximation (3.282) the correcting divergence term of the liquid compression will be expressed by $-\boldsymbol{q}_l \gamma^l \rho_0^l g \cdot \nabla h^l \approx -\boldsymbol{q}_l \frac{S_o^l}{\varepsilon} \cdot \nabla h^l$.

3.10.5 Basic Model Equations of Single Liquid Phase Variable-Density Darcy-Type Flow, Mass and Heat Transport in Variably Saturated Porous Media: Second Level Reduction

Applying the above assumptions and derivations of Sects. 3.10.1–3.10.4 to the equations of the first level reduction as listed in Table 3.5, we can now summarize the governing balance laws with their related constitutive relations in Table 3.7 as the basic model equations of second level reduction, which are formulated in the D−dimensional Euclidean space $\Re^D$ $(D = 1, 2, 3)$. Because we assume that only one dynamic fluid phase, the liquid phase l, is present, we can omit the index l in the symbols for the sake of simplicity. Only the solid phase needs to be further identified by the index s. Typical adsorption relations for the adsorption function φ_k, the retardation factor $\Re_k$ and the derivative term of retardation $\acute{\Re}_k$ are listed in Table 3.8, which are derived in detail in Chap. 5.

3.10.6 Basic Model Equations of Variable-Density Darcy-Type Flow, Mass and Heat Transport in Groundwater: Third Level Reduction

The equations of Table 3.7 can be simplified for flow, mass and heat transport in groundwater, the fully saturated porous medium. In this case

- The saturation is set to $s = 1$

and the following system of equations results which is summarized in Table 3.9.

Table 3.7 Summarized balance laws and constitutive relations (CR) of single liquid phase variable-density Darcy-type flow, mass and heat transport in variably saturated porous media as *second level* model reduction. It forms a system of $2 + N^* + D$ equations[a] to solve the (2) variables h (or ψ)[b] and T, the (N^*) variables C_k of species k (or C_m^s of species m)[c] in the fluid phase l and in the solid phase s, respectively, and the (D) variables $\boldsymbol{q}$. Alternative convective forms are given in angle brackets, alternative variable formulation of the Darcy law is given in round brackets.

Type	Equations
$\mathcal{M}^l$	$s\,S_o \dfrac{\partial h}{\partial t} + \varepsilon \dfrac{\partial s}{\partial t} + \nabla \cdot \boldsymbol{q} = Q + Q_{\text{EOB}}$
$\mathcal{M}_k^l + \mathcal{M}_k^s$	$\dfrac{\partial}{\partial t}(\varepsilon s \mathfrak{R}_k C_k) + \nabla \cdot (\boldsymbol{q} C_k) + \nabla \cdot \boldsymbol{j}_k + \varepsilon s \vartheta_k \mathfrak{R}_k C_k = \tilde{R}_k$ $\left\langle \varepsilon s \acute{\mathfrak{R}}_k \dfrac{\partial C_k}{\partial t} + \boldsymbol{q} \cdot \nabla C_k + \nabla \cdot \boldsymbol{j}_k + \varepsilon s \vartheta_k \mathfrak{R}_k C_k = \tilde{R}_k - C_k Q \right\rangle$
$\mathcal{M}_m^s$	$\dfrac{\partial}{\partial t}(\varepsilon_s C_m^s) = \varepsilon_s (r_m^s + R_m^s) \qquad m \neq k$ $\left\langle \varepsilon_s \dfrac{\partial C_m^s}{\partial t} = \varepsilon_s (r_m^s + R_m^s) \right\rangle$
$\mathcal{V}^l$	$\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})$ $\Big(\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot \big[\nabla \psi + (1 + \chi)\boldsymbol{e}\big]\Big)$
$\mathcal{E}^l + \mathcal{E}^s$	$\dfrac{\partial}{\partial t}\Big[\big(\varepsilon s \rho c + (1-\varepsilon)\rho^s c^s\big)(T - T_0)\Big] + \nabla \cdot (\rho c \boldsymbol{q}(T - T_0)) - \nabla \cdot (\boldsymbol{\Lambda} \cdot \nabla T) = H_e$ $\left\langle \big(\varepsilon s \rho c + (1-\varepsilon)\rho^s c^s\big) \dfrac{\partial T}{\partial t} + \rho c \boldsymbol{q} \cdot \nabla T - \nabla \cdot (\boldsymbol{\Lambda} \cdot \nabla T) = H_e - \rho c (T - T_0) Q \right\rangle$
CR	$\chi = \sum_{k=1}^{N^l-1} \beta_{c_k}(C_k - C_{k0}) - \beta(T)(T - T_0)$ $Q_{\text{EOB}} = -\boldsymbol{q} \cdot \Big(\frac{S_o}{\varepsilon}\nabla h + \sum_{k=1}^{N^l-1} \beta_{c_k} \nabla C_k - \beta^* \nabla T\Big) - \varepsilon s \Big(\sum_{k=1}^{N^l-1} \beta_{c_k} \frac{\partial C_k}{\partial t} - \beta^* \frac{\partial T}{\partial t}\Big)$ $\beta^* = \begin{cases} \beta & \text{constant} \\ \dfrac{\beta(T) + \frac{\partial \beta(T)}{\partial T}(T - T_0)}{1 + \sum_{k=1}^{N^l-1} \beta_{c_k}(C_k - C_{k0}) - \beta(T)(T - T_0)} & \text{variable}^d \end{cases}$ $\beta_{c_k} = \frac{\alpha_k}{C_{ks} - C_{k0}}$ $S_o = \rho_0 g(\varepsilon \gamma + \upsilon)$ $s = s(\psi)$ $k_r = k_r(s)$ $\boldsymbol{j}_k(\Im_H \lVert \boldsymbol{j}_k \rVert + 1) = -\boldsymbol{D}_k \cdot \nabla C_k$ $\boldsymbol{D}_k = \varepsilon s D_k \boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}}$ $\boldsymbol{D}_{\text{mech}} = \beta_T \lVert \boldsymbol{q} \rVert \boldsymbol{\delta} + (\beta_L - \beta_T) \frac{\boldsymbol{q} \otimes \boldsymbol{q}}{\lVert \boldsymbol{q} \rVert}$ $\mathfrak{R}_k = 1 + \big(\frac{1-\varepsilon}{\varepsilon}\big)\varphi_k$ $\acute{\mathfrak{R}}_k = 1 + \big(\frac{1-\varepsilon}{\varepsilon}\big) \frac{\partial(\varphi_k C_k)}{\partial C_k}$ $\varphi_k = \varphi_k(C_k)$ $\boldsymbol{\Lambda} = \big[\varepsilon s \Lambda + (1-\varepsilon)\Lambda^s\big]\boldsymbol{\delta} + \rho c \boldsymbol{D}_{\text{mech}}$ $f_\mu = \mu_0 / \mu(\frac{C_k}{\rho}, T)$ $H_e = \rho H + \rho^s H_s$

[a] Liquid phase index l is omitted for the sake of simplicity.

[b] $h = \psi + x_j$, the saturation s appears as a secondary variable which can be computed via the capillary pressure head relations $s(\psi)$ described in Appendix D.

[c] Species k can occur both in the liquid phase l and the solid phase s, however, species $m \neq k$ only occurs in the solid phase s.

[d] The function $\beta(T)$ and its derivative $\partial\beta(T)/\partial T$ are formulated in Appendix C: (C.8) and (C.10), respectively.

Table 3.8 Typical adsorption function φ_k, retardation factor $\Re_k$ and derivative term of retardation $\acute{\Re}_k$. The parameter κ_k is the Henry sorptivity coefficient[a], $b_k^{\dagger}$ and $b_k^{\ddagger}$ are the coefficient and exponent, respectively, and $k_k^{\dagger}$ and $k_k^{\ddagger}$ are the coefficients. Derivation is given in Chap. 5

Type of isotherm	φ_k	$\Re_k$	$\acute{\Re}_k$
Henry	κ_k	$1+\left(\frac{1-\varepsilon}{\varepsilon}\right)\kappa_k$	$1+\left(\frac{1-\varepsilon}{\varepsilon}\right)\kappa_k$
Freundlich	$b_k^{\dagger}\,C_k^{b_k^{\ddagger}-1}$	$1+\left(\frac{1-\varepsilon}{\varepsilon}\right)b_k^{\dagger}\,C_k^{b_k^{\ddagger}-1}$	$1+\left(\frac{1-\varepsilon}{\varepsilon}\right)b_k^{\dagger}b_k^{\ddagger}\,C_k^{b_k^{\ddagger}-1}$
Langmuir	$\frac{k_k^{\dagger}}{1+k_k^{\ddagger}\,C_k}$	$1+\left(\frac{1-\varepsilon}{\varepsilon}\right)\frac{k_k^{\dagger}}{1+k_k^{\ddagger}\,C_k}$	$1+\left(\frac{1-\varepsilon}{\varepsilon}\right)\frac{k_k^{\dagger}}{(1+k_k^{\ddagger}\,C_k)^2}$

[a] The *Henry isotherm* with a retardation factor of $\Re_k = 1 + \left(\frac{1-\varepsilon}{\varepsilon}\right)\kappa_k$ is often expressed by the *distribution coefficient* K_k^d in the form [39]: $\Re_k = 1 + \left(\frac{1-\varepsilon}{\varepsilon}\right)\rho^s K_k^d$ with $K_k^d = \kappa_k/\rho^s$. Another alternative definition of the distribution coefficient can sometimes be found as $\Re_k = 1+\left(\frac{\rho_s K_k^d}{\varepsilon}\right)$, where $\rho_s = (1-\varepsilon)\rho^s$ is the bulk density of porous media (mass dry media per total volume).

3.10.7 Basic Model Equations of Vertically Averaged Flow, Mass and Heat Transport in Unconfined and Confined Aquifers: Specific Case of Third Level Reduction

Flow, mass and heat transport, which are essentially horizontal in an aquifer, can be vertically averaged as described in Sect. 3.5. 2D depth-integrated balance equations result as described in Sect. 3.7.8 for which constitutive relations have to be added similar to those as developed for the full 3D problems above. In doing this, the following simplifications for the 2D, vertically averaged, essentially horizontal flow and transport processes in aquifers hold:

- The aquifer forms a layer of a saturated porous medium of thickness $B = B(x_1, x_2, t)$. While the bottom of the layer is considered stationary, on top the saturated zone is bounded by a possibly moving phreatic surface, so that as shown in Fig. 3.13:

$$\begin{aligned} B(x_1,x_2,t) &= h(x_1,x_2,t) - f^B(x_1,x_2) \text{ for unconfined condition} \\ B(x_1,x_2) \quad &= f^T(x_1,x_2) - f^B(x_1,x_2) \text{ for confined condition} \end{aligned} \tag{3.283}$$

  where $h = h(x_1, x_2, t)$ is the hydraulic head (3.260), $f^T(x_1, x_2)$ and $f^B(x_1, x_2)$ are the top and bottom bounding surfaces, respectively.
- The coordinate direction of integration x_j (commonly $x_j = x_3 = z$) coincides with direction of gravity, i.e., $\nabla x_j = \boldsymbol{e}$. Accordingly, gravitational effects on liquid density disappear. (Extensions will be treated in Sect. 11.9).

The boundaries of the aquifer on top and bottom, respectively, can be expressed by their surface functions, cf. (2.112)

Table 3.9 Summarized balance laws and constitutive relations (CR) of variable-density Darcy-type flow, mass and heat transport in groundwater as *third level* model reduction. It forms a system of $2 + N^* + D$ equations[a] to solve the (2) variables h and T, the (N^*) variables C_k of species k (or C_m^s of species m)[b] in the fluid phase l and in the solid phase s, respectively, and the (D) variables $\boldsymbol{q}$. Alternative convective forms are given in angle brackets.

Type	Equations
$\mathcal{M}^l$	$S_o \dfrac{\partial h}{\partial t} + \nabla \cdot \boldsymbol{q} = Q + Q_{\text{EOB}}$
$\mathcal{M}_k^l + \mathcal{M}_k^s$	$\dfrac{\partial}{\partial t}(\varepsilon \mathfrak{R}_k C_k) + \nabla \cdot (\boldsymbol{q} C_k) + \nabla \cdot \boldsymbol{j}_k + \varepsilon \vartheta_k \mathfrak{R}_k C_k = \tilde{R}_k$ $\Big\langle \varepsilon \acute{\mathfrak{R}}_k \dfrac{\partial C_k}{\partial t} + \boldsymbol{q} \cdot \nabla C_k + \nabla \cdot \boldsymbol{j}_k + \varepsilon \vartheta_k \mathfrak{R}_k C_k = \tilde{R}_k - C_k Q \Big\rangle$
$\mathcal{M}_m^s$	$\dfrac{\partial}{\partial t}(\varepsilon_s C_m^s) = \varepsilon_s (r_m^s + R_m^s) \qquad m \neq k$ $\Big\langle \varepsilon_s \dfrac{\partial C_m^s}{\partial t} = \varepsilon_s (r_m^s + R_m^s) \Big\rangle$
$\mathcal{V}^l$	$\boldsymbol{q} = -\boldsymbol{K} f_\mu \cdot \big(\nabla h + \chi \boldsymbol{e}\big)$
$\mathcal{E}^l + \mathcal{E}^s$	$\dfrac{\partial}{\partial t}\Big[\big(\varepsilon \rho c + (1-\varepsilon)\rho^s c^s\big)(T - T_0)\Big] + \nabla \cdot (\rho c \boldsymbol{q}(T - T_0)) - \nabla \cdot (\boldsymbol{\Lambda} \cdot \nabla T) = H_e$ $\Big\langle \big(\varepsilon \rho c + (1-\varepsilon)\rho^s c^s\big) \dfrac{\partial T}{\partial t} + \rho c \boldsymbol{q} \cdot \nabla T - \nabla \cdot (\boldsymbol{\Lambda} \cdot \nabla T) = H_e - \rho c (T - T_0) Q \Big\rangle$
CR	$\begin{aligned} \chi &= \textstyle\sum_{k=1}^{N^l-1} \beta_{c_k}(C_k - C_{k0}) - \beta(T)(T - T_0) \\ Q_{\text{EOB}} &= -\boldsymbol{q} \cdot \Big(\tfrac{S_o}{\varepsilon} \nabla h + \textstyle\sum_{k=1}^{N^l-1} \beta_{c_k} \nabla C_k - \beta^* \nabla T \Big) - \varepsilon \Big(\textstyle\sum_{k=1}^{N^l-1} \beta_{c_k} \tfrac{\partial C_k}{\partial t} - \beta^* \tfrac{\partial T}{\partial t} \Big) \\ \beta^* &= \begin{cases} \beta & \text{constant} \\ \dfrac{\beta(T) + \frac{\partial \beta(T)}{\partial T}(T - T_0)}{1 + \sum_{k=1}^{N^l-1} \beta_{c_k}(C_k - C_{k0}) - \beta(T)(T - T_0)} & \text{variable}^{c} \end{cases} \\ \beta_{c_k} &= \tfrac{\alpha_k}{C_{ks} - C_{k0}} \\ S_o &= \rho_0 g (\varepsilon \gamma + \upsilon) \\ \boldsymbol{j}_k (\Im_H \lVert \boldsymbol{j}_k \rVert + 1) &= -\boldsymbol{D}_k \cdot \nabla C_k \\ \boldsymbol{D}_k &= \varepsilon D_k \boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}} \\ \boldsymbol{D}_{\text{mech}} &= \beta_T \lVert \boldsymbol{q} \rVert \boldsymbol{\delta} + (\beta_L - \beta_T) \tfrac{\boldsymbol{q} \otimes \boldsymbol{q}}{\lVert \boldsymbol{q} \rVert} \\ \mathfrak{R}_k &= 1 + \big(\tfrac{1-\varepsilon}{\varepsilon}\big) \varphi_k \\ \acute{\mathfrak{R}}_k &= 1 + \big(\tfrac{1-\varepsilon}{\varepsilon}\big) \tfrac{\partial(\varphi_k C_k)}{\partial C_k} \\ \varphi_k &= \varphi_k(C_k) \\ \boldsymbol{\Lambda} &= \big[\varepsilon \Lambda + (1-\varepsilon)\Lambda^s\big] \boldsymbol{\delta} + \rho c \boldsymbol{D}_{\text{mech}} \\ f_\mu &= \mu_0 / \mu(\tfrac{C_k}{\rho}, T) \\ H_e &= \rho H + \rho^s H_s \end{aligned}$

[a] Liquid phase index l is omitted for the sake of simplicity.

[b] Species k can occur both in the liquid phase l and the solid phase s, however, species $m \neq k$ only occurs in the solid phase s.

[c] The function $\beta(T)$ and its derivative $\partial \beta(T)/\partial T$ are formulated in Appendix C: (C.8) and (C.10), respectively.

$$\begin{aligned} F^T &= F^T(x_1, x_2, x_3, t) = x_3 - h(x_1, x_2, t) = 0 \text{ unconfined} \\ F^T &= F^T(x_1, x_2, x_3) \quad = x_3 - f^T(x_1, x_2) = 0 \text{ confined} \\ F^B &= F^B(x_1, x_2, x_3) \quad = x_3 - f^B(x_1, x_2) = 0 \text{ unconfined/confined} \end{aligned} \tag{3.284}$$

For unconfined conditions the top boundary moves with the velocity $\boldsymbol{w}$ and in accordance with (2.113)–(2.115) it is

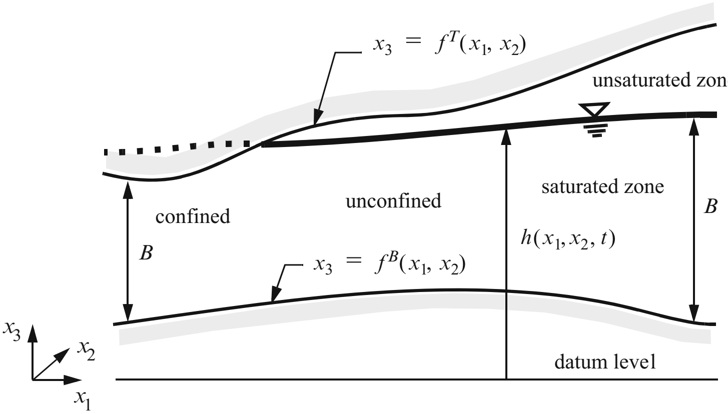


Fig. 3.13 Unconfined and confined conditions in an aquifer (vertical cross section $x_j = x_3$)

$$\begin{aligned} &\frac{\partial F^T}{\partial t} + \boldsymbol{w} \cdot \nabla F^T = 0 \quad \text{or} \\ &\frac{\partial h}{\partial t} - \boldsymbol{w} \cdot \nabla (x_3 - h) = 0 \end{aligned} \tag{3.285}$$

with the outward-pointing unit normal vector to the surface $F^T = 0$

$$\boldsymbol{n} = \frac{\nabla F^T}{\|\nabla F^T\|} \tag{3.286}$$

and the normal component of the moving surface with $F^T = x_3 - h = 0$

$$\boldsymbol{w} \cdot \boldsymbol{n} = -\frac{\partial F^T/\partial t}{\|\nabla F^T\|} = \frac{\partial h}{\partial t} \tag{3.287}$$

where we have assumed that $\|\nabla F^T\| = \|\nabla(x_3 - h)\| \approx \|\nabla x_3\| = 1$, i.e., the water table is approximately horizontal. For the stationary boundaries F^T in the case of confined aquifer and F^B we have

$$\boldsymbol{w} \cdot \boldsymbol{n} = 0 \quad \left.\frac{\partial F^T}{\partial t}\right|_{\text{confined}} = \frac{\partial F^B}{\partial t} = 0 \tag{3.288}$$

Now, let us consider the vertically averaged mass balance equation (3.76) written in the form:

$$\frac{\partial}{\partial t}(B\varepsilon\rho) + \nabla \cdot (B\varepsilon\rho\boldsymbol{v}) = B\varepsilon\rho(Q + Q_{\text{ex}}) \tag{3.289}$$

We can replace the external mass supply Q_{ex} by the interface relation (3.75) of the upper phreatic surface specified for mass

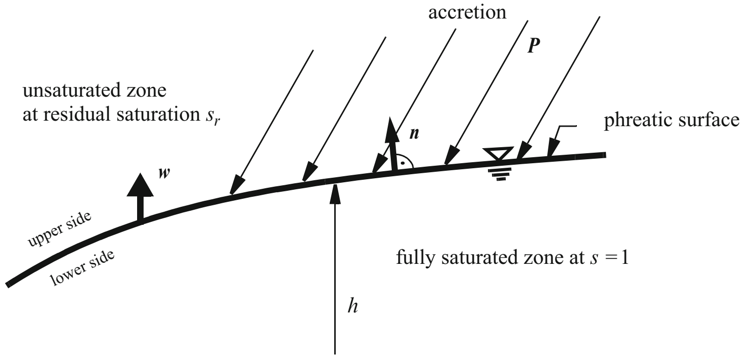


Fig. 3.14 Phreatic surface with accretion

$$B\varepsilon\rho Q_{\text{ex}} = -\varepsilon\rho(\boldsymbol{v} - \boldsymbol{w}) \cdot \boldsymbol{n} \tag{3.290}$$

where $\boldsymbol{w}$ designates the macroscopic surface velocity and $\boldsymbol{n}$ is the outward unit normal vector to the moving surface F^T. The phreatic surface separates the fully saturated zone from the unsaturated zone, where we assume that the interface is sharp. It forms the water table with $h = x_3$. For the unsaturated zone we assume that the liquid in the void space is at the residual (irreducible) saturation s_r. On the upper side of the phreatic surface we take into account the possibility of accretion $\boldsymbol{P}$, e.g., from precipitation, as depicted in Fig. 3.14.

The mass balance at the phreatic surface requires that the mass flux through the lower side of the interface at the saturated zone is equal to the mass flux through the upper side of the interface at the unsaturated zone, viz.,

$$\varepsilon(\boldsymbol{v} - \boldsymbol{w})\big|_{\text{sat}} \cdot \boldsymbol{n} - \varepsilon(\boldsymbol{v} - \boldsymbol{w})\big|_{\text{unsat}} \cdot \boldsymbol{n} = 0 \tag{3.291}$$

The accretion is $\boldsymbol{P} = \varepsilon\boldsymbol{v}\big|_{\text{unsat}}$. For a vertically downward-oriented accretion we use

$$\boldsymbol{P} = -P\,\nabla x_3 \tag{3.292}$$

where P corresponds to the rate of infiltration or groundwater recharge. Using (3.287) with $\varepsilon|_{\text{unsat}} = \varepsilon s_r$ and $\boldsymbol{w}\big|_{\text{sat}} = \boldsymbol{w}\big|_{\text{unsat}}$ we find

$$\varepsilon\boldsymbol{w}\big|_{\text{unsat}} \cdot \boldsymbol{n} = \varepsilon s_r \frac{\partial h}{\partial t} \tag{3.293}$$

Then, with (3.291), (3.292), and (3.293) the interface BC reads

$$\begin{aligned} B\varepsilon Q_{\text{ex}} &= -\varepsilon(\boldsymbol{v} - \boldsymbol{w})\big|_{\text{sat}} \cdot \boldsymbol{n} \\ = -\varepsilon(\boldsymbol{v} - \boldsymbol{w})\big|_{\text{unsat}} \cdot \boldsymbol{n} &= \underbrace{-\boldsymbol{P} \cdot \boldsymbol{n}}_{P} + \varepsilon s_r \frac{\partial h}{\partial t} \end{aligned} \tag{3.294}$$

and the flux BC of a phreatic (free) surface results

$$\underbrace{\varepsilon \boldsymbol{v}}_{\boldsymbol{q}} \cdot \boldsymbol{n} = \underbrace{\varepsilon(1-s_r)}_{\varepsilon_e} \frac{\partial h}{\partial t} - P \tag{3.295}$$

where

$$\varepsilon_e = \varepsilon(1-s_r) \tag{3.296}$$

is referred to as the *specific yield* (also called storativity or *drainable and fillable porosity*) of a phreatic aquifer and $\boldsymbol{q} \cdot \boldsymbol{n}$ is the positive outward normal flux of liquid leaving the saturated zone through the phreatic surface.

Using (3.294) the mass balance equation (3.289) for the unconfined aquifer can be written as

$$\frac{B\varepsilon}{\rho}\frac{\partial \rho}{\partial t} + B\frac{\partial \varepsilon}{\partial t} + \varepsilon\frac{\partial B}{\partial t} + \nabla \cdot (B\boldsymbol{q}) = B\varepsilon Q + P \; + \; \varepsilon s_r \frac{\partial h}{\partial t} \tag{3.297}$$

Since $B = h - f^B$ (3.283) the vertically averaged mass balance equation (3.297) for an unconfined aquifer finally takes the form:

$$(S_o B + \varepsilon_e)\frac{\partial h}{\partial t} + \nabla \cdot (B\boldsymbol{q}) = B\varepsilon Q + P \tag{3.298}$$

where the derivations of $\partial\rho/\partial t$ and $\partial\varepsilon/\partial t$ have been developed in the same manner as described in Sect. 3.10.4. To simplify the notation we shall designate depth-integrated quantities by an overline and define

$$\begin{aligned} \bar{\boldsymbol{q}} &= B\,\boldsymbol{q} \\ \bar{S}_o &= B\,S_o \\ \bar{Q} &= B\varepsilon Q + P \end{aligned} \tag{3.299}$$

so that (3.298) can be written as

$$(\bar{S}_o + \varepsilon_e)\frac{\partial h}{\partial t} + \nabla \cdot \bar{\boldsymbol{q}} = \bar{Q} \tag{3.300}$$

The remaining vertically averaged balance equations (3.77)–(3.82) for species mass, momentum and energy can now be similarly developed, where the same principles for the constitutive relations are applied as described in Sect. 3.10.6 for the fully saturated porous medium (groundwater). The resulting model equations of vertically averaged flow, mass and heat transport in an unconfined aquifer are summarized in Table 3.10.

Under confined aquifer conditions the boundary surfaces F^T and F^B are assumed stationary, so that $\partial B/\partial t = 0$ and $B\varepsilon Q_{\mathrm{ex}} = P$ and the mass balance

Table 3.10 Summarized balance laws and constitutive relations (CR) of vertically averaged flow, mass and heat transport in an unconfined aquifer forming a system of $4 + N^*$ equations[a] to solve the (2) variables h and T, the (N^*) variables C_k of species k (or C_m^s of species m)[b] in the fluid phase l and in the solid phase s, respectively, and the (2) variables $\bar{\boldsymbol{q}}$. Alternative convective forms are given in angle brackets.

Type	Equations
$\mathcal{M}^l$	$(\bar{S}_o + \varepsilon_e)\dfrac{\partial h}{\partial t} + \nabla \cdot \bar{\boldsymbol{q}} = \bar{Q}$
$\mathcal{M}_k^l + \mathcal{M}_k^s$	$\dfrac{\partial}{\partial t}(\varepsilon \bar{\mathfrak{R}}_k C_k) + \nabla \cdot (\bar{\boldsymbol{q}} C_k) + \nabla \cdot \bar{\boldsymbol{j}}_k + \varepsilon \vartheta_k \bar{\mathfrak{R}}_k C_k = \bar{\tilde{R}}_k$ $\left\langle \varepsilon \bar{\acute{\mathfrak{R}}}_k \dfrac{\partial C_k}{\partial t} + \bar{\boldsymbol{q}} \cdot \nabla C_k + \nabla \cdot \bar{\boldsymbol{j}}_k + \varepsilon \vartheta_k \bar{\mathfrak{R}}_k C_k = \bar{\tilde{R}}_k - C_k \bar{Q} \right\rangle$
$\mathcal{M}_m^s$	$\dfrac{\partial}{\partial t}(\varepsilon_s B\, C_m^s) = \varepsilon_s B(r_m^s + R_m^s) \qquad m \neq k$ $\left\langle \varepsilon_s B \dfrac{\partial C_m^s}{\partial t} = \varepsilon_s B(r_m^s + R_m^s) \right\rangle$
$\mathcal{V}^l$	$\bar{\boldsymbol{q}} = -B\, \boldsymbol{K} f_\mu \cdot \nabla h$
$\mathcal{E}^l + \mathcal{E}^s$	$\dfrac{\partial}{\partial t}\Big[B\big(\varepsilon \rho c + (1-\varepsilon)\rho^s c^s\big)(T - T_0)\Big] + \nabla \cdot (\rho c \bar{\boldsymbol{q}}(T - T_0)) - \nabla \cdot (\bar{\boldsymbol{\Lambda}} \cdot \nabla T) = \bar{H}_e$ $\left\langle B\big(\varepsilon \rho c + (1-\varepsilon)\rho^s c^s\big)\dfrac{\partial T}{\partial t} + \rho c \bar{\boldsymbol{q}} \cdot \nabla T - \nabla \cdot (\bar{\boldsymbol{\Lambda}} \cdot \nabla T) = \bar{H}_e - \rho c (T - T_0)\bar{Q} \right\rangle$
CR	$B = h - f^B$ $\bar{S}_o = \rho_0 g B(\varepsilon \gamma + \upsilon)$ $\bar{\boldsymbol{j}}_k(\bar{\Im}_H \Vert \bar{\boldsymbol{j}}_k \Vert + 1) = -\bar{\boldsymbol{D}}_k \cdot \nabla C_k$ $\bar{\boldsymbol{D}}_k = \varepsilon B D_k \boldsymbol{\delta} + \bar{\boldsymbol{D}}_{\mathrm{mech}}$ $\bar{\boldsymbol{D}}_{\mathrm{mech}} = \beta_T \Vert \bar{\boldsymbol{q}} \Vert \boldsymbol{\delta} + (\beta_L - \beta_T)\frac{\bar{\boldsymbol{q}} \otimes \bar{\boldsymbol{q}}}{\Vert \bar{\boldsymbol{q}} \Vert}$ $\bar{\mathfrak{R}}_k = B\big[1 + \big(\frac{1-\varepsilon}{\varepsilon}\big)\varphi_k\big]$ $\bar{\acute{\mathfrak{R}}}_k = B\big[1 + \big(\frac{1-\varepsilon}{\varepsilon}\big)\frac{\partial(\varphi_k C_k)}{\partial C_k}\big]$ $\varphi_k = \varphi_k(C_k)$ $\bar{\boldsymbol{\Lambda}} = B\big[\varepsilon \Lambda + (1-\varepsilon)\Lambda^s\big]\boldsymbol{\delta} + \rho c \bar{\boldsymbol{D}}_{\mathrm{mech}}$ $f_\mu = \mu_0/\mu(\frac{C_k}{\rho}, T)$ $\bar{H}_e = B(\rho H + \rho^s H_s)$ $\bar{\tilde{R}}_k = B\tilde{R}_k$

[a] The gradient operator ∇ is only 2D.

[b] Species k can occur both in the liquid phase l and the solid phase s, however, species $m \neq k$ only occurs in the solid phase s.

equation (3.300) reduces to the simple form

$$\bar{S}_o \frac{\partial h}{\partial t} + \nabla \cdot \bar{\boldsymbol{q}} = \bar{Q} \tag{3.301}$$

Usually, for confined aquifers the product of hydraulic conductivity $\boldsymbol{K}$ (3.263) and aquifer thickness $B = f^T - f^B$ (3.283) is combined in the *tensor of transmissivity* $\boldsymbol{T}$ defined as

Table 3.11 Summarized balance laws and constitutive relations (CR) of vertically averaged flow, mass and heat transport in a confined aquifer forming a system of $4 + N^*$ equations[a] to solve the (2) variables h and T, the (N^*) variables C_k of species k (or C_m^s of species m)[b] in the fluid phase l and in the solid phase s, respectively, and the (2) variables $\bar{\boldsymbol{q}}$. Alternative convective forms are given in angle brackets.

Type	Equations
$\mathcal{M}^l$	$\bar{S}_o \dfrac{\partial h}{\partial t} + \nabla \cdot \bar{\boldsymbol{q}} = \bar{Q}$
$\mathcal{M}_k^l + \mathcal{M}_k^s$	$\dfrac{\partial}{\partial t}(\varepsilon \bar{\mathfrak{R}}_k C_k) + \nabla \cdot (\bar{\boldsymbol{q}} C_k) + \nabla \cdot \bar{\boldsymbol{j}}_k + \varepsilon \vartheta_k \bar{\mathfrak{R}}_k C_k = \bar{\tilde{R}}_k$ $\left\langle \varepsilon \bar{\acute{\mathfrak{R}}}_k \dfrac{\partial C_k}{\partial t} + \bar{\boldsymbol{q}} \cdot \nabla C_k + \nabla \cdot \bar{\boldsymbol{j}}_k + \varepsilon \vartheta_k \bar{\mathfrak{R}}_k C_k = \bar{\tilde{R}}_k - C_k \bar{Q} \right\rangle$
$\mathcal{M}_m^s$	$\dfrac{\partial}{\partial t}(\varepsilon_s B\, C_m^s) = \varepsilon_s B(r_m^s + R_m^s) \qquad m \neq k$ $\left\langle \varepsilon_s B \dfrac{\partial C_m^s}{\partial t} = \varepsilon_s B(r_m^s + R_m^s) \right\rangle$
$\mathcal{V}^l$	$\bar{\boldsymbol{q}} = -\boldsymbol{T} f_\mu \cdot \nabla h$
$\mathcal{E}^l + \mathcal{E}^s$	$\dfrac{\partial}{\partial t}\Big[B\big(\varepsilon \rho c + (1-\varepsilon)\rho^s c^s\big)(T - T_0)\Big] + \nabla \cdot (\rho c \bar{\boldsymbol{q}}(T - T_0)) - \nabla \cdot (\bar{\boldsymbol{\Lambda}} \cdot \nabla T) = \bar{H}_e$ $\left\langle B\big(\varepsilon \rho c + (1-\varepsilon)\rho^s c^s\big) \dfrac{\partial T}{\partial t} + \rho c \bar{\boldsymbol{q}} \cdot \nabla T - \nabla \cdot (\bar{\boldsymbol{\Lambda}} \cdot \nabla T) = \bar{H}_e - \rho c (T - T_0) \bar{Q} \right\rangle$
CR	$B = f^T - f^B$ $\bar{S}_o = \rho_0 g B(\varepsilon \gamma + \upsilon)$ $\bar{\boldsymbol{j}}_k (\bar{\mathfrak{S}}_H \lVert \bar{\boldsymbol{j}}_k \rVert + 1) = -\bar{\boldsymbol{D}}_k \cdot \nabla C_k$ $\bar{\boldsymbol{D}}_k = \varepsilon B D_k \boldsymbol{\delta} + \bar{\boldsymbol{D}}_{\text{mech}}$ $\bar{\boldsymbol{D}}_{\text{mech}} = \beta_T \lVert \bar{\boldsymbol{q}} \rVert \boldsymbol{\delta} + (\beta_L - \beta_T) \frac{\bar{\boldsymbol{q}} \otimes \bar{\boldsymbol{q}}}{\lVert \bar{\boldsymbol{q}} \rVert}$ $\bar{\mathfrak{R}}_k = B\big[1 + \big(\frac{1-\varepsilon}{\varepsilon}\big)\varphi_k\big]$ $\bar{\acute{\mathfrak{R}}}_k = B\big[1 + \big(\frac{1-\varepsilon}{\varepsilon}\big) \frac{\partial(\varphi_k C_k)}{\partial C_k}\big]$ $\varphi_k = \varphi_k(C_k)$ $\bar{\boldsymbol{\Lambda}} = B\big[\varepsilon \Lambda + (1-\varepsilon)\Lambda^s\big]\boldsymbol{\delta} + \rho c \bar{\boldsymbol{D}}_{\text{mech}}$ $f_\mu = \mu_0 / \mu(\frac{C_k}{\rho}, T)$ $\bar{H}_e = B(\rho H + \rho^s H_s)$ $\bar{\tilde{R}}_k = B \tilde{R}_k$

[a] The gradient operator ∇ is only 2D.

[b] Species k can occur both in the liquid phase l and the solid phase s, however, species $m \neq k$ only occurs in the solid phase s.

$$\boldsymbol{T} = B\boldsymbol{K} = \frac{\boldsymbol{k} B \rho_0 g}{\mu_0} \tag{3.302}$$

which represents an aquifer property measured as the flow rate per unit width through the entire aquifer thickness. The concept of transmissivity is only applicable to vertically averaged, essentially horizontal flow in confined aquifers. In Table 3.11 we summarize the governing model equations of vertically averaged flow, mass and heat transport in confined aquifers.

3.11 Standard Model Equations for Solving Flow, Mass and Heat Transport in Porous Media

The equation systems derived in Sect. 3.10 provide the general physical modeling basis for solving 3D and 2D (including axisymmetric and vertically averaged) variable-density flow, multispecies (chemically reactive) mass and heat transport processes in variably saturated porous media. They are summarized in Tables 3.7, 3.9, 3.10, and 3.11 for the variably saturated porous medium, for groundwater, for 2D unconfined and confined aquifers, respectively. In general, the equations are nonlinearly coupled due to density effects, dependencies by variable saturation (or presence of phreatic surface), chemical reactions, non-Fickian mass flux and viscosity effects. Four problem classes are distinguished:

1. *Flow:* Solving the flow equations in a separate manner, there are no density and viscosity effects.
2. *Flow + mass:* Solving flow and mass transport, which can be coupled by density, viscosity, chemical reaction and non-Fickian mass flux.
3. *Flow + heat:* Solving flow and heat transport, which can be coupled by density and viscosity.
4. *Flow + mass + heat:* This represents the most complex model for simultaneous solution of flow, mass and heat transport, which can be coupled by density, viscosity, chemical reaction and non-Fickian mass flux. If the non-isothermal mass transport is related to salinity, the processes are often termed as *thermohaline*. Since, in general, mass and heat have different diffusivities new phenomena can result for this problem class termed as *double-diffusive convection* (DDC).

With respect to the temporal dependency of the governing flow and transport equations we can choose three time classes:

(i) *Transient flow/transient transport:* Both flow and (mass/heat) transport are simulated in their fully temporal dependency as formulated in the basic equations of Tables 3.7, 3.9, 3.10, and 3.11.

(ii) *Steady-state flow/transient transport:* The flow process is considered stationary, i.e., $\partial h/\partial t = 0$, while (mass/heat) transport remains fully transient. This exceptional case is useful if the temporal, often short-term variations of flow are negligible in comparison to the temporal, often long-term variations of mass and/or heat transport.

(iii) *Steady-state flow/stationary transport:* For both flow and (mass/heat) transport only steady-state solutions are searched, i.e., $\partial h/\partial t = 0$, $\partial C_k/\partial t = 0$ and $\partial T/\partial t = 0$. We note, however, under nonlinear conditions in general, a unique steady-state solution must not exist.

It is obvious that the Darcy law of momentum conservation, written in the general form

$$\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot \big(\nabla h + \chi \boldsymbol{e}\big) \tag{3.303}$$

Table 3.12 Primary and secondary variables of standard model equations

Type	Equations set	Primary variables[a]	Secondary variables[b]
Variably saturated media	Table 3.7	h, C_k, T	$s, \boldsymbol{q}, (\psi, p)$
Fully saturated media (groundwater)	Table 3.9	h, C_k, T	$\boldsymbol{q}, (p)$
2D unconfined aquifer[c]	Table 3.10	h, C_k, T	$\boldsymbol{q}$
2D confined aquifer[c]	Table 3.11	h, C_k, T	$\boldsymbol{q}$

[a] Number of variables is $2 + N^*$: 1 for flow, N^* for mass and 1 for heat, independent of problem dimension D

[b] Other secondary variables could be mass and heat fluxes $\boldsymbol{j}_k$ and $\boldsymbol{j}_T$, respectively

[c] Essentially horizontal, vertically averaged equations

is very well suited for substituting the Darcy velocity $\boldsymbol{q}$ in the mass conservation

$$s\, S_o \frac{\partial h}{\partial t} + \varepsilon \frac{\partial s}{\partial t} + \nabla \cdot \boldsymbol{q} = Q + Q_{\mathrm{EOB}} \tag{3.304}$$

to obtain the flow equation

$$s\, S_o \frac{\partial h}{\partial t} + \varepsilon \frac{\partial s}{\partial t} - \nabla \cdot \left[k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e}) \right] = Q + Q_{\mathrm{EOB}} \tag{3.305}$$

which represents a generalized form of the so-called *Richards' equation* named after L.A. Richards [440], who firstly derived and published such a type of flow equation for unsaturated porous media in 1931. The advantage is that only *one* primary variable h (or ψ) remains to be solved for the flow problem, while the Darcy velocity $\boldsymbol{q}$ appears now as a secondary variable, which can be easily solved from (3.303) with known h. Accordingly, the solution is reduced to only 1 flow equation, N^* mass transport equations and 1 heat transport equation of scalar primary variables h, C_k and T, respectively, as summarized in Table 3.12.

Chapter 4
Discrete Feature

4.1 Discrete-Feature Conceptual Model

In Sect. 3.2 the REV concept was introduced for porous media, where for an effective continuum a certain scale exists at which the geometric properties of the void space and individual heterogeneities are smoothed out due to the process of spatial averaging. Even for fractured media of rock masses the averaging technique can also be applicable, provided the scale of the void space and heterogeneity of both pores and fractures forms an equivalent continuum of an overlapping REV. In a fracture domain the scale of heterogeneity is the distance between fractures. If the domain is sufficiently large, a REV can be found and the fractured porous media can be treated as a single continuum for which the modeling equations of Chap. 3 are valid. We note, however, that the scale of REV cannot be too large and must be significantly smaller than the scale of gross inhomogeneity and scale of problem domain as required by (3.1).

In some cases, flow, mass and heat transport can tend to be dominated by a limited number of discrete pathways formed by fractures and other discrete features clearly represented by different scales for which an equivalent single (overlapping) REV does not exist anymore and different continua, one for the porous medium and one for the discrete feature, must be applied. Porous media and discrete features are then treated separately and have to be coupled through the macroscopic interfaces. The term *discrete features* is used here as a generalization of all those geometric representations of a lower spatial dimension having commonly a significant fluid conductance in comparison to the porous medium. Accordingly, discrete features do not mean only fractures, they encompass various natural and engineered characteristics such as listed in Table 4.1 and illustrated in Fig. 4.1.

The central component in modeling discrete features is that their geometries and properties are explicitly incorporated. A discrete feature can often be used as a domain, where one or two spatial dimensions are much smaller than the length of interest, so that they can be represented by 2D or even only 1D geometries. For instance, a discrete feature in form of a fracture is characterized by an

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_4,

Table 4.1 Typical discrete features

Dimension	Application
1D (plane)	Channel, mine stope, tunnel, river
1D (tubular)	Pumping well, abandoned well, vug, borehole, shaft, karstic conduit, drift, tunnel, cavity, drain
2D	Fracture, fault, overland flow, runoff

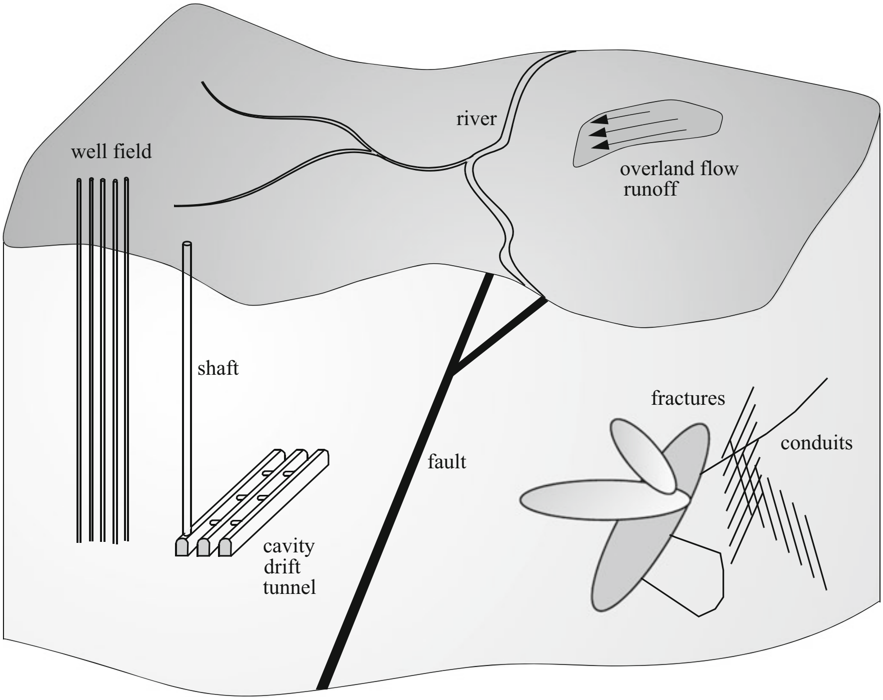


Fig. 4.1 Subsurface and surface hydro-geosystem structure containing discrete features

aperture which is very small and the flow in a fracture takes place essentially parallel to the fracture's axis. On the other hand, the diameter of a borehole and its cross-sectional flow variations are usually very small compared to the length, so that flow and transport through a borehole can be represented by a 1D process dominated along the borehole axis. Apparently, the range of applications and the dimension of the discrete features require a unified approach possessing different laws of fluid motion, both pure fluid and porous-medium flows, problems with and without free surface in plane (1D, 2D) and tubular (1D) geometries.

Table 4.2 Balance laws

Quantity	$\rho\psi$	$\boldsymbol{j}$	ρF
Mass			
barycentric	ρ	$\mathbf{0}$	ρQ
species	C_k	$\boldsymbol{j}_k$	r_k
Momentum	$\rho\boldsymbol{v}$	$\boldsymbol{\sigma}$	$\rho\boldsymbol{g}$
Energy	$\rho(E + \frac{1}{2}v^2)$	$\boldsymbol{j}_T + \boldsymbol{\sigma}\cdot\boldsymbol{v}$	$\rho(H + \boldsymbol{g}\cdot\boldsymbol{v})$

4.2 Fundamental Relations

The flow, mass and heat transport equations were derived in Chap. 3 for a porous medium which is represented by a multiphase system. They consist of both D−dimensional and depth-integrated 2D equations. From the mathematical point of view these equations represent generalized formulations from which the basic equations valid for discrete features can easily be degenerated by a formal reduction of the number of dimension and number of phases. For instance, the pure fluid flow in a fracture is nothing more than a single-fluid phase process with a porosity $\varepsilon = 1$ and saturation $s = 1$. Furthermore, 2D depth-integrated fracture flow and transport is in a full analogy to the vertically averaged processes in aquifers as already described in Chap. 3.

4.2.1 General Balance Statement

The conservation of mass, momentum and energy is described by the balance statement (3.34) written in the form

$$\frac{\partial(\rho\psi)}{\partial t} + \nabla\cdot(\rho\psi\ \boldsymbol{v}) + \nabla\cdot\boldsymbol{j} = \rho F \tag{4.1}$$

conserving the (extensive) quantity $(\rho\psi)$. Individual balance laws for $(\rho\psi)$, $\boldsymbol{j}$ and ρF are summarized in Table 4.2.

4.2.2 Forms of Balance Equations

According to the typical applications for discrete features indicated above, four forms of the governing balance equation (4.1) can be distinguished:

- *Form A*: pure fluid balance law
- *Form B*: vertically integrated pure fluid balance law
- *Form C*: porous-medium balance law
- *Form D*: vertically integrated porous-medium balance law

The *form A* is already represented by (4.1). A vertical integration of (4.1) over a depth (thickness, aperture) B can be rigorously performed as described in Sect. 3.5 leading to the *form B*:

$$\frac{\partial(B\rho\psi)}{\partial t} + \nabla \cdot (B\rho\psi\boldsymbol{v}) + \nabla \cdot (B\boldsymbol{j}) = B\rho F + j_\psi^T - j_\psi^B \tag{4.2}$$

with the new exchange terms of the quantity ψ at the top and bottom boundaries, respectively

$$\begin{aligned} j_\psi^T &= \frac{1}{dS}\int_{dS^T} \boldsymbol{n}^T \cdot [\boldsymbol{j} + \rho\psi(\boldsymbol{w} - \boldsymbol{v})]da \\ j_\psi^B &= \frac{1}{dS}\int_{dS^B} \boldsymbol{n}^B \cdot [\boldsymbol{j} + \rho\psi(\boldsymbol{w} - \boldsymbol{v})]da \end{aligned} \tag{4.3}$$

Note that the balance quantities of (4.2) are now averaged over the depth B and the gradient operator ∇ is only 2D. The transformation of the balance equation (4.1) to a porous medium is performed by spatial averaging procedures referred to the REV as described in Sect. 3.3. It finally yields the *form C* of the basic balance statement written for a single liquid phase of saturation s

$$\frac{\partial(\varepsilon s\rho\psi)}{\partial t} + \nabla \cdot (\varepsilon s\rho\psi\boldsymbol{v}) + \nabla \cdot (\varepsilon s\boldsymbol{j}) = \varepsilon s\rho F + j_\psi^I \tag{4.4}$$

where ε corresponds to the porosity (void space) of the porous medium. In (4.4) an exchange term at the liquid-solid interface appears

$$j_\psi^I = \frac{1}{dS}\int_{dS^I} \boldsymbol{n}^I \cdot [\boldsymbol{j} + \rho\psi(\boldsymbol{w} - \boldsymbol{v})]da \tag{4.5}$$

Note that the balance quantities of the porous-medium conservation equation (4.4) are averaged over the REV volume. Finally, the porous-medium equation (4.4) can also be vertically integrated over the depth B, which yields *form D* of the basic balance statement as

$$\frac{\partial(B\varepsilon s\rho\psi)}{\partial t} + \nabla \cdot (B\varepsilon s\rho\psi\boldsymbol{v}) + \nabla \cdot (B\varepsilon s\boldsymbol{j}) = B\varepsilon s\rho F + Bj_\psi^I + j_\psi^T - j_\psi^B \tag{4.6}$$

It is obvious, from the mathematical point of view the balance statement (4.6) of form D represents the most general formulation which comprises all other forms if we specify the porosity ε as

$$\varepsilon \begin{cases} \equiv 1 \\ < 1 \end{cases} \text{for} \quad \begin{array}{l} \text{pure liquid flow} \\ \text{porous-medium flow} \end{array} \tag{4.7}$$

the saturation s as

$$s \begin{cases} \equiv 1 \\ \leq 1 \end{cases} \text{for} \quad \begin{matrix} \text{pure liquid flow} \\ \text{porous-medium flow} \end{matrix} \tag{4.8}$$

the depth B as

$$B \begin{cases} \equiv 1 \\ > 0 \end{cases} \text{for} \quad \begin{matrix} \text{non-integrated form} \\ \text{vertically integrated form} \end{matrix} \tag{4.9}$$

the interface exchange term j_ψ^I as

$$j_\psi^I \begin{cases} \equiv 0 \\ \neq 0 \end{cases} \text{for} \quad \begin{matrix} \text{pure liquid flow} \\ \text{porous-medium flow} \end{matrix} \tag{4.10}$$

and the top and bottom exchange terms j_ψ^T, j_ψ^B as

$$(j_\psi^T, j_\psi^B) \begin{cases} \equiv 0 \\ \neq 0 \end{cases} \text{for} \quad \begin{matrix} \text{non-integrated form} \\ \text{vertically integrated form} \end{matrix} \tag{4.11}$$

4.2.3 Hydraulic Radius

The hydraulic radius is defined as the flow cross-sectional area divided by the wetted perimeter

$$r_{\text{hydr}} = \frac{\text{flow area}}{\text{wetted perimeter}} \tag{4.12}$$

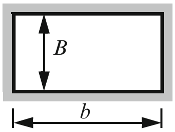
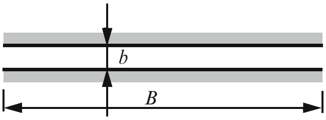
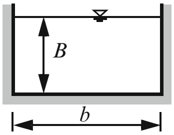
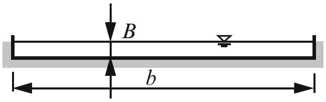
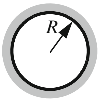
Table 4.3 lists the hydraulic radii for interesting cases.

4.2.4 Free Surface Condition

A free 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, cf. (2.113)

$$\frac{\partial F}{\partial t} + \boldsymbol{w} \cdot \nabla F = 0 \tag{4.13}$$

Table 4.3 Hydraulic radii for different applications

Case	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>20 B), plane 	$\dfrac{B}{1+2B/b} \approx B$
(e)	submerged circular cross - section 	$\dfrac{\pi R^2}{2\pi R} = \dfrac{R}{2}$

where $\boldsymbol{w}$ is the velocity of the interface. The outward unit vector normal to F is defined as

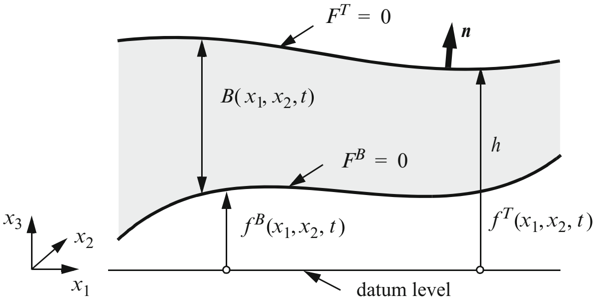
$$\boldsymbol{n} = \frac{\nabla F}{\|\nabla F\|} \tag{4.14}$$

and accordingly

$$\boldsymbol{w} \cdot \boldsymbol{n} = -\frac{\partial F/\partial t}{\|\nabla F\|} \tag{4.15}$$

where $\|\nabla F\|$ denotes the magnitude of the vector ∇F. For the vertical integration over the thickness B, similar to Sect. 3.10.7, we can express the geometries of the top and bottom surfaces in the forms (Fig. 4.2), cf. (3.284)

Fig. 4.2 Surface conditions



$$
\begin{aligned}
F^T &= F^T(\boldsymbol{x}, t) = x_3 - f^T(x_1, x_2, t) = 0 \\
F^B &= F^B(\boldsymbol{x}, t) = x_3 - f^B(x_1, x_2, t) = 0
\end{aligned}
\tag{4.16}
$$

and

$$
B = B(\boldsymbol{x}, t) = f^T(x_1, x_2, t) - f^B(x_1, x_2, t) \tag{4.17}
$$

For a free surface the top elevation $x_3 = f^T(x_1, x_2, t)$ is identical to the hydraulic head $h = h(x_1, x_2, t)$. Accordingly, the thickness is given by

$$
B = B(\boldsymbol{x}, t) = h - f^B \tag{4.18}
$$

4.2.5 Viscous Stresses on Surfaces

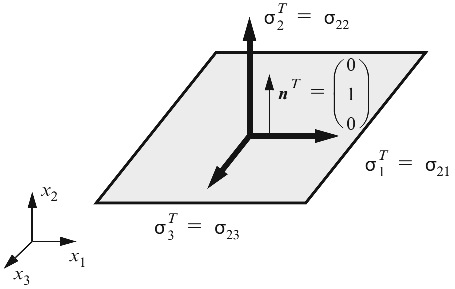
The viscous stresses on top and bottom surfaces result from exchange relationships (4.3) if replacing the general flux vector $\boldsymbol{j}$ by the viscous stress tensor of liquid $\boldsymbol{\sigma}$ (cf. Table 4.2), viz.,

$$
\boldsymbol{\sigma}^{TB} = \frac{1}{dS} \int_{dS^{TB}} \boldsymbol{n}^{TB} \cdot [\boldsymbol{\sigma} + \rho \boldsymbol{v}(\boldsymbol{w} - \boldsymbol{v})] da \tag{4.19}
$$

where $\boldsymbol{\sigma}^{TB}$ stands for the viscous stress on the top and bottom surface with normal $\boldsymbol{n}^{TB}$. It represents a surface force per unit area depending on the orientation of the surface [409]. For instance, let us consider the stress components on a planar top surface as illustrated in Fig. 4.3. Assuming additionally a rigid and impermeable surface ($\boldsymbol{w} = \boldsymbol{0}$, $\boldsymbol{n}^T \cdot \boldsymbol{v} = 0$) with a constant stress property on the unit area dS, the surface stress is explicitly given by

$$
\boldsymbol{\sigma}^T = \boldsymbol{n}^T \cdot \boldsymbol{\sigma} \tag{4.20}
$$

Fig. 4.3 Surface forces related to the components of the viscous stress tensor $\boldsymbol{\sigma}$



$\sigma_2^T = \sigma_{22}$
$\boldsymbol{n}^T = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$
$\sigma_1^T = \sigma_{21}$
$\sigma_3^T = \sigma_{23}$
x_2
x_1
x_3

With $\boldsymbol{n}^T = (0\ 1\ 0)^T$ the stress components become

$$\begin{aligned} \sigma_1^T &= 0\,\sigma_{11} + 1\,\sigma_{21} + 0\,\sigma_{31} = \sigma_{21} \\ \sigma_2^T &= 0\,\sigma_{12} + 1\,\sigma_{22} + 0\,\sigma_{32} = \sigma_{22} \\ \sigma_3^T &= 0\,\sigma_{13} + 1\,\sigma_{23} + 0\,\sigma_{33} = \sigma_{23} \end{aligned} \tag{4.21}$$

4.3 Basic Balance Laws

4.3.1 Liquid Mass Conservation

The liquid mass conservation is described by specifying (4.6) with Table 4.2 as

$$\frac{\partial}{\partial t}(B\varepsilon s\rho) + \nabla \cdot (B\varepsilon s\rho \boldsymbol{v}) = B\varepsilon s\rho Q + BQ^I + Q^T - Q^B \tag{4.22}$$

or by introducing the Darcy velocity $\boldsymbol{q} = \varepsilon s(\boldsymbol{v} - \boldsymbol{v}^s)$, cf. (3.240),

$$\frac{\partial}{\partial t}(B\varepsilon s\rho) + \nabla \cdot (B\rho\, \boldsymbol{q}) + \nabla \cdot (B\rho\varepsilon s \boldsymbol{v}^s) = B\varepsilon s\rho Q + BQ^I + Q^T - Q^B \tag{4.23}$$

where Q^I, Q^T and Q^B are the interfacial, top and bottom exchange terms, respectively. Using the constitutive relations for liquid and medium compressibility, (3.197), (3.274), (3.246), (3.281), as well as assuming unconfined conditions for $B = h - f^B$, (3.283), and the approximation for deformable media (3.243), (3.245), we find for the terms of (4.23)

$$\begin{aligned} \frac{\partial}{\partial t}(B\varepsilon s\rho) + \nabla \cdot (B\rho\varepsilon s \boldsymbol{v}^s) &= B\varepsilon s\frac{\partial \rho}{\partial t} + \rho\Big(\frac{Bs}{1-\varepsilon}\frac{\partial \varepsilon}{\partial t} + \varepsilon s\frac{\partial B}{\partial t} + B\varepsilon\frac{\partial s}{\partial t}\Big) \\ &= \rho\Big[\big(Bs\varepsilon\rho_0 g\gamma + Bs\upsilon\rho_0 g + \varepsilon s\big)\frac{\partial h}{\partial t} + B\varepsilon\frac{\partial s}{\partial t}\Big] \\ &= \rho\Big[\big(BsS_o + \varepsilon s\big)\frac{\partial h}{\partial t} + B\varepsilon\frac{\partial s}{\partial t}\Big] \end{aligned} \tag{4.24}$$

with the specific storage coefficient

$$S_o = \rho_0 g(\varepsilon\gamma + \upsilon) \tag{4.25}$$

Taking the surface relationships introduced in Sect. 3.10.7 with (3.290) to (3.296) we obtain for the top exchange terms of (4.23):

$$\begin{aligned} Q^T &= -\varepsilon s \rho(\boldsymbol{v} - \boldsymbol{w}) \cdot \boldsymbol{n} = -\varepsilon s_r \rho(\boldsymbol{v} - \boldsymbol{w})|_{\text{unsat}} \cdot \boldsymbol{n} \\ &= \rho P + \rho\varepsilon s_r \frac{\partial h}{\partial t} \end{aligned} \tag{4.26}$$

where P is the recharge. Finally, by using (4.24) and (4.26) the liquid mass conservation (4.23) obtains the form

$$\big[BsS_o + \varepsilon(s - s_r)\big]\frac{\partial h}{\partial t} + B\varepsilon\frac{\partial s}{\partial t} + \nabla \cdot (B\boldsymbol{q}) = \bar{Q} \tag{4.27}$$

with a generalized source/sink term $\bar{Q} = B\varepsilon s Q + BQ^I + P - Q^B$. In (4.26) the OB approximation (cf. Sect. 3.10.3) is applied, where density effects in the mass balance equation are neglected. We note that the specification of ε and s in (4.27) has to be in accordance with the problem classes to be used. For instance, for a pure liquid flow it is $\varepsilon = 1$ and $s = 1$ and the mass conservation equation reads

$$\begin{aligned} BS_o\frac{\partial h}{\partial t} + \nabla \cdot (B\boldsymbol{q}) &= \bar{Q} \\ S_o &= \rho_0 g\gamma \\ \boldsymbol{q} &= \boldsymbol{v} \end{aligned} \tag{4.28}$$

On the other hand, a fully saturated porous medium with $s = 1$, the mass balance equation (4.27) simplifies

$$\big(BS_o + \varepsilon_e\big)\frac{\partial h}{\partial t} + \nabla \cdot (B\boldsymbol{q}) = \bar{Q} \tag{4.29}$$

where $\varepsilon_e = \varepsilon(1 - s_r)$ is the specific yield (3.296). It is important to note that for a variably saturated porous medium the free surface (phreatic) condition in (4.27) is not applicable. In this case the contribution $\varepsilon(s - s_r)$ in the storage term of (4.27) must vanish, cf. Sect. 3.10.7.

4.3.2 *Liquid Momentum Conservation*

The liquid momentum conservation is specified from (4.6) with Table 4.2 as

$$\frac{\partial}{\partial t}(B\varepsilon s\rho\boldsymbol{v}) + \nabla\cdot(B\varepsilon s\rho\boldsymbol{v}\boldsymbol{v}) = -B\varepsilon s\nabla p + \nabla\cdot(B\varepsilon s\boldsymbol{\tau}) + \\ B\varepsilon s\rho\boldsymbol{g} + B\boldsymbol{f}_\tau + \boldsymbol{\sigma}^T - \boldsymbol{\sigma}^B \qquad (4.30)$$

where the liquid stress tensor $\boldsymbol{\sigma}$ is split into the equilibrium part (pressure p) and non-equilibrium (deviatory) stress $\boldsymbol{\tau}$ according to (3.114). In (4.30) $\boldsymbol{f}_\tau$, $\boldsymbol{\sigma}^T$ and $\boldsymbol{\sigma}^B$ represent the deviatoric interfacial, the top and bottom momentum exchange vectors, respectively. Notice, in (4.30) the momentum exchange term $\boldsymbol{f}_\tau$ vanishes for pure liquid motion and the terms $\boldsymbol{\sigma}^T$ and $\boldsymbol{\sigma}^B$ are dropped if the equation is not vertically integrated.

In the following we assume the Newton's viscosity law (3.142) (including the Stokes' assumption (3.141)) which is written in the form

$$\boldsymbol{\tau} = \tfrac{2}{3}\mu(\boldsymbol{\delta}:\boldsymbol{d})\boldsymbol{\delta} - 2\mu\boldsymbol{d} \qquad (4.31)$$

with the strain-rate tensor

$$\boldsymbol{d} = \tfrac{1}{2}\left[\nabla\boldsymbol{v} + (\nabla\boldsymbol{v})^T\right] \qquad (4.32)$$

where μ is the dynamic viscosity of the liquid. For an incompressible liquid with a divergenceless (solenoidal) velocity $\nabla\cdot\boldsymbol{v} = 0$, the momentum conservation equation (4.30) leads to the well-known *Navier-Stokes equation* written in a generalized form as

$$B\varepsilon s\rho\left[\frac{\partial\boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v}\right] = -B\varepsilon s(\nabla p - \rho\boldsymbol{g}) + B\mu\nabla^2(\varepsilon s\boldsymbol{v}) + \\ B\boldsymbol{f}_\tau + \boldsymbol{\sigma}^T - \boldsymbol{\sigma}^B \qquad (4.33)$$

from where more specific forms will be derived as follows.

4.3.2.1 Darcy Flow in Porous Media

Commonly, in a porous medium the velocity $\boldsymbol{v}$ is sufficiently small so that the Reynolds number Re_p (3.254) based on a typical pore diameter persists in a range $\text{Re}_p < 1\ldots10$. As a consequence, the inertial terms in the momentum equation (4.33) can be neglected

$$\frac{\partial\boldsymbol{v}}{\partial t} \approx \mathbf{0} \qquad (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} \approx \mathbf{0} \qquad (4.34)$$

and the momentum equation for porous media written in its non-integrated form with $B \equiv 1$ and $\boldsymbol{\sigma}^T = \boldsymbol{\sigma}^B = \mathbf{0}$ yields

$$\varepsilon s(\nabla p - \rho\boldsymbol{g}) = \boldsymbol{f}_\tau + \mu\nabla^2(\varepsilon s\boldsymbol{v}) \qquad (4.35)$$

Furthermore, the viscous shear stresses of the liquid appearing in the Brinkman term $\mu\nabla^2(\varepsilon s \boldsymbol{v})$ can be usually neglected in comparison to the drag term of momentum exchange $\boldsymbol{f}_\tau$ as discussed in Sect. 3.9.2. The interfacial drag term of momentum exchange $\boldsymbol{f}_\tau$ can be derived as a linear relationship of the form, cf. (3.154):

$$\boldsymbol{f}_\tau = -\varepsilon s \mu (\boldsymbol{k}^l)^{-1} \cdot \boldsymbol{q} \tag{4.36}$$

where the intrinsic permeability $\boldsymbol{k}^l$ represents an inverse friction tensor due to the viscous drag at the liquid-solid interfaces of the porous medium, which is expressed as $\boldsymbol{k}^l = k_r \boldsymbol{k}$, cf. (3.224). Finally, the momentum equation (4.35) reduces to the well-known *Darcy equation* written in its pressure formulation

$$\boldsymbol{q} = -\frac{k_r \boldsymbol{k}}{\mu} \cdot (\nabla p - \rho \boldsymbol{g}) \tag{4.37}$$

or in its equivalent hydraulic head formulation[1]

$$\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e}) \tag{4.38}$$

where $h = p/(\rho_0 g) + x_j$, $\boldsymbol{K} = \boldsymbol{k}\rho_0 g/\mu_0$, $f_\mu = \mu_0/\mu$, $\chi = (\rho - \rho_0)/\rho_0$ and $\boldsymbol{e} = -\boldsymbol{g}/\|\boldsymbol{g}\|$ (see also Sect. 3.10.2). Equations (4.37) and (4.38) are valid for flow in a discrete feature filled by a porous medium.

4.3.2.2 Plane and Axisymmetric Parallel (Hagen-Poiseuille) Flow

A pure liquid flow is called parallel when inertial terms of the Navier-Stokes equation (4.33) vanish. That means, a liquid particle is subjected to zero acceleration, accordingly, it moves in pure translation with constant velocity $\boldsymbol{v}$. 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 laminar flow fields occur between two parallel plates or in a circular tube as depicted in Fig. 4.4. They are termed as *Hagen-Poiseuille flow* named after the German engineer G. Hagen (1839) and the French physician J. Poiseuille (1840) who first studied independently this type of flow [409].

[1] With the hydraulic head $h = p/(\rho_0 g) + x_j = \psi + x_j$, (3.260), it is $p = \rho_0 g(h - x_j)$ and yields:

$$\nabla p - \rho \boldsymbol{g} = \nabla p + \rho g \boldsymbol{e} = \rho_0 g \Big(\nabla\psi + \nabla x_j + \frac{\rho - \rho_0}{\rho_0} \Big) = \rho_0 g (\nabla h + \chi \boldsymbol{e}).$$

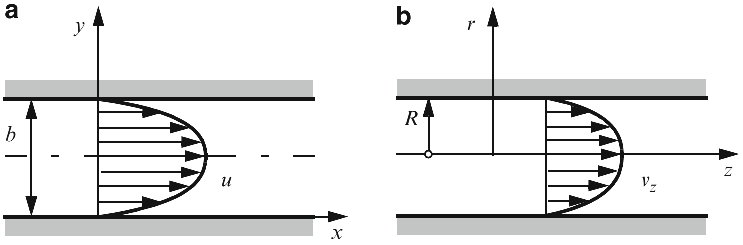


Fig. 4.4 (**a**) 2D plane and (**b**) axisymmetric Hagen-Poiseuille flow

For a 2D parallel laminar flow (Fig. 4.4a) we have

$$\boldsymbol{v} = \begin{pmatrix} u \\ v \\ w \end{pmatrix} \qquad u = u(y) \qquad v = w = 0 \tag{4.39}$$

and the momentum equation (4.33) in the $x-$direction becomes (note that we consider the pure liquid case $\varepsilon s \equiv 1$ without vertical integration $B \equiv 1$):

$$\frac{dp}{dx} - \rho g_x = \mu \frac{d^2 u}{dy^2} \tag{4.40}$$

Integrating (4.40) with the BC $u(0) = u(b) = 0$, where b is the aperture (Fig. 4.4a), it yields

$$u = -\frac{1}{2\mu}\Big(\frac{dp}{dx} - \rho g_x\Big) y(b-y) \tag{4.41}$$

and we obtain the average velocity $\bar{u}$ in the aperture b as

$$\bar{u} = \frac{1}{b}\int_{y=0}^{b} u\, dy = -\frac{b^2}{12\mu}\Big(\frac{dp}{dx} - \rho g_x\Big) \tag{4.42}$$

and the discharge Q (per unit width) as

$$Q = \bar{u}\, b = -\frac{b^3}{12\mu}\Big(\frac{dp}{dx} - \rho g_x\Big) \tag{4.43}$$

which is called the *cubic law* of the Hagen-Poiseuille flow. The relationship (4.42) can be expressed by the hydraulic radius r_{hydr} if replacing the dimension $b/2$ for the slit flow according to Table 4.3, case (b):

$$\bar{u} = -\frac{r_{\text{hydr}}^2}{3\mu}\Big(\frac{dp}{dx} - \rho g_x\Big) \tag{4.44}$$

Similarly, for the axisymmetric flow in a circular tube (Fig. 4.4b) with

$$\boldsymbol{v} = \begin{pmatrix} v_r \\ v_\phi \\ v_z \end{pmatrix} \qquad v_z = v_z(r) \qquad v_r = v_\phi = 0 \tag{4.45}$$

we solve the momentum equation (4.33) in the z–direction (see (2.73) for the second-order derivative operation in cylindrical coordinates):

$$\frac{dp}{dz} - \rho g_z = \frac{\mu}{r}\left[\frac{\partial}{\partial r}\Big(r\frac{\partial v_z}{\partial r}\Big)\right] \tag{4.46}$$

With $dv_z/dr = 0$ at $r = 0$ and $v_z(R) = 0$ (Fig. 4.4b) the integration of (4.46) gives

$$v_z = -\frac{1}{4\mu}\Big(\frac{dp}{dz} - \rho g_z\Big)(R^2 - r^2) \tag{4.47}$$

Then, the average velocity for the Hagen-Poiseuille flow in a circular tube becomes

$$\bar{v}_z = \frac{1}{\pi R^2}\int_{\phi=0}^{2\pi}\int_{r=0}^{R} v_z\, r\, dr d\phi = -\frac{R^2}{8\mu}\Big(\frac{dp}{dz} - \rho g_z\Big) \tag{4.48}$$

and the discharge through the tube is

$$Q = \pi R^2\, \bar{v}_z = -\frac{\pi R^4}{8\mu}\Big(\frac{dp}{dz} - \rho g_z\Big) \tag{4.49}$$

The relationship (4.48) can be expressed by the hydraulic radius r_{hydr} if replacing the dimension $R/2$ for the tube flow according to Table 4.3, case (e):

$$\bar{v}_z = -\frac{r_{\text{hydr}}^2}{2\mu}\Big(\frac{dp}{dz} - \rho g_z\Big) \tag{4.50}$$

As seen the Hagen-Poiseuille's laws of laminar liquid motion for plane flow (4.42) and axisymmetric flow (4.48) represent linear relationships with respect to the pressure gradient and gravity ($\nabla p - \rho \boldsymbol{g}$). Instead of p we can formulate the relationships with the hydraulic head h and find the following generalized equation of 'diffusive flux-type' for the Hagen-Poiseuille flow:

$$\begin{aligned} &\boldsymbol{v} = -\boldsymbol{K} f_\mu(\nabla h + \chi \boldsymbol{e}) \\ &\boldsymbol{K} = \frac{r_{\text{hydr}}^2\, \rho_0 g}{a\, \mu_0}\, \boldsymbol{\delta} \quad \text{with} \quad \begin{cases} r_{\text{hydr}} = b/2,\ a = 3 \text{ for plane flow} \\ r_{\text{hydr}} = R/2,\ a = 2 \text{ for axisymmetric flow} \end{cases} \end{aligned} \tag{4.51}$$

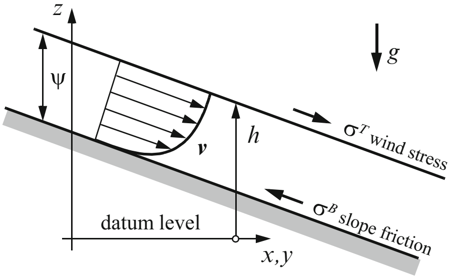


Fig. 4.5 Open channel flow

4.3.2.3 Laws of Liquid Motion for Overland and Channel Flow

Basically, the pure liquid motion for overland and channel flow is described by the vertically integrated Navier-Stokes equation (4.33) according to

$$B\rho\Big[\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\nabla)\boldsymbol{v}\Big] = -B(\nabla p - \rho \boldsymbol{g}) + B\mu\nabla^2\boldsymbol{v} + \boldsymbol{\sigma}^T - \boldsymbol{\sigma}^B \tag{4.52}$$

which is a formulation of the well-known *De Saint-Venant equation* [73, 74, 245, 483]. Over a wide range of practical overland and channel flow (Fig. 4.5) at low-to-moderate velocity/flow regimes the inertial terms in the governing momentum balance equation (4.52) can be ignored compared to the gravitational terms, friction and pressure effects. Furthermore, the interior viscous effects can be neglected over the shear stress effects at the surfaces [73, 74]. Assuming that,

$$\frac{\partial \boldsymbol{v}}{\partial t} \approx \boldsymbol{0} \qquad (\boldsymbol{v}\cdot\nabla)\boldsymbol{v} \approx \boldsymbol{0} \qquad \mu\nabla^2\boldsymbol{v} \approx \boldsymbol{0} \tag{4.53}$$

the momentum equation (4.52) reduces to

$$B(\nabla p - \rho\boldsymbol{g}) - \boldsymbol{\sigma}^T + \boldsymbol{\sigma}^B = \boldsymbol{0} \tag{4.54}$$

The shear effect $\boldsymbol{\sigma}^T$ on the top (free) surface can be caused by wind stress. However, for the present applications influences by wind stress will be neglected:

$$\boldsymbol{\sigma}^T \approx \boldsymbol{0} \tag{4.55}$$

On the other hand, the shear effects at the bottom surface $\boldsymbol{\sigma}^B$ represent the dominant friction forces and can usually expressed by a friction slope relationship of the form:

$$\begin{aligned} \boldsymbol{\sigma}^B &= \rho_0 g B \boldsymbol{S}_f \\ \boldsymbol{S}_f &= \frac{\|\boldsymbol{v}\|\boldsymbol{v}}{\tau^2\, r_{\text{hydr}}^{\delta}} \end{aligned} \tag{4.56}$$

Table 4.4 Various friction laws

Law	τ	δ	$\boldsymbol{S}_f$
Newton-Taylor	$\sqrt{\frac{g}{f_N}}$	1	$\frac{f_N \Vert\boldsymbol{v}\Vert\boldsymbol{v}}{g\, r_{\text{hydr}}}$
Darcy-Weisbach	$\sqrt{\frac{8g}{f_D}}$	1	$\frac{f_D \Vert\boldsymbol{v}\Vert\boldsymbol{v}}{8g\, r_{\text{hydr}}}$
Chezy	C	1	$\frac{\Vert\boldsymbol{v}\Vert\boldsymbol{v}}{C^2\, r_{\text{hydr}}}$
Manning-Strickler	M	$4/3$	$\frac{\Vert\boldsymbol{v}\Vert\boldsymbol{v}}{M^2\, r_{\text{hydr}}^{4/3}}$

where $\boldsymbol{S}_f$ is the vector of friction slopes at channel bottom, $\Vert\boldsymbol{v}\Vert = \sqrt{\boldsymbol{v}\cdot\boldsymbol{v}}$, τ is a general friction factor and $\delta \geq 1$ is a constant. Specifications of τ and δ provide different friction laws as listed in Table 4.4 for isotropic roughness coefficients.

Inserting (4.56) with (4.55) into (4.54) the following momentum equation results

$$(\nabla p - \rho \boldsymbol{g}) + \rho_0 g \boldsymbol{S}_f = \boldsymbol{0} \tag{4.57}$$

Instead of using the pressure p as primary variable the hydraulic head h or the local pressure head ψ are alternative formulations of (4.57), viz.,

$$\nabla h + \boldsymbol{S}_f + \chi \boldsymbol{e} = \boldsymbol{0} \tag{4.58}$$

or

$$\nabla \psi + \boldsymbol{S}_f + (1+\chi)\boldsymbol{e} = \boldsymbol{0} \tag{4.59}$$

where the buoyancy coefficient χ comprises liquid density effects. Equation (4.58) can be used to derive a *diffusion-type flow equation* [195]. Since, exemplified for 2D

$$\Vert\boldsymbol{v}\Vert^2 = u^2 + v^2 = \tau^2\, r_{\text{hydr}}^{\delta} \sqrt{S_{fx}^2 + S_{fy}^2} \tag{4.60}$$

and using (4.56)

$$S_{fx} = \frac{\sqrt{u^2+v^2}}{\tau^2\, r_{\text{hydr}}^{\delta}}\, u \qquad S_{fy} = \frac{\sqrt{u^2+v^2}}{\tau^2\, r_{\text{hydr}}^{\delta}}\, v \tag{4.61}$$

we find with (4.58): $S_{fx} = -(\partial h/\partial x + \chi e_x)$, $S_{fy} = -(\partial h/\partial y + \chi e_y)$

$$\begin{aligned} u &= -\frac{\tau\, r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{(\frac{\partial h}{\partial x})^2 + (\frac{\partial h}{\partial y})^2}} \left(\frac{\partial h}{\partial x} + \chi e_x\right) \\ v &= -\frac{\tau\, r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{(\frac{\partial h}{\partial x})^2 + (\frac{\partial h}{\partial y})^2}} \left(\frac{\partial h}{\partial y} + \chi e_y\right) \end{aligned} \tag{4.62}$$

which can be concisely written in a generalized ‘diffusive flux-type’ form:

$$\begin{aligned} \boldsymbol{v} &= -\boldsymbol{K} \cdot (\nabla h + \chi \boldsymbol{e}) \\ \boldsymbol{K} &= \frac{\tau\, r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\|\nabla h\|^2}}\, \boldsymbol{\delta} \end{aligned} \tag{4.63}$$

It can be easily shown that the velocity $\boldsymbol{v}$ in (4.63) tends to zero if the gradient ∇h vanishes, provided that $\chi = 0$:

$$\lim_{\nabla h \to \mathbf{0}} \boldsymbol{v} = -\lim_{\nabla h \to \mathbf{0}} \frac{\tau\, r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\|\nabla h\|^2}}\, \boldsymbol{\delta} \cdot \nabla h = \mathbf{0} \tag{4.64}$$

4.3.3 Species Mass Conservation

The mass conservation of species k is specified from (4.6) with Table 4.2 as

$$\frac{\partial (B \varepsilon s C_k)}{\partial t} + \nabla \cdot (B \varepsilon s C_k \boldsymbol{v}) + \nabla \cdot (B \varepsilon s \boldsymbol{j}_k) = B \varepsilon s \bar{r}_k \tag{4.65}$$

which can be employed for all interesting mass transport problems when specifying εs and B appropriately. Note that the reaction term $\bar{r}_k$ also includes both interfacial and surfacial mass transfer conditions. In analogy to the developments done in Sect. 3.9.2 the reaction term $\bar{r}_k$ can be split into a first-order reaction rate and a production term, respectively,

$$\bar{r}_k = -\vartheta_k C_k + \tilde{R}_k \tag{4.66}$$

The species mass flux $\boldsymbol{j}_k$ is expressed by the *Fick’s law* (3.183) written in the form

$$\begin{aligned} \boldsymbol{j}_k &= -\boldsymbol{D}_k \cdot \nabla C_k \\ \boldsymbol{D}_k &= D_k \boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}} \end{aligned} \tag{4.67}$$

The hydrodynamic dispersion tensor $\boldsymbol{D}_k$ consists of the molecular diffusion part $D_k \boldsymbol{\delta}$ and the mechanical dispersion part $\boldsymbol{D}_{\text{mech}}$. In a porous medium $\boldsymbol{D}_{\text{mech}}$ is usually described by the Scheidegger-Bear dispersion model (3.182) as

$$\boldsymbol{D}_{\text{mech}} = \beta_T \|\boldsymbol{v}\| \boldsymbol{\delta} + (\beta_L - \beta_T) \frac{\boldsymbol{v} \otimes \boldsymbol{v}}{\|\boldsymbol{v}\|} \tag{4.68}$$

In a pure liquid flow there is a large variety for $\boldsymbol{D}_{\text{mech}}$ in dependence on laminar and turbulent flow conditions. For instance, in a liquid-filled tube under laminar flow conditions $\boldsymbol{D}_{\text{mech}}$ can be estimated by Taylor's analysis [508]:

$$\boldsymbol{D}_{(k)\text{mech}} = \left(\frac{R^2\|\boldsymbol{v}\|}{48D_k}\right)\frac{\boldsymbol{v}\otimes\boldsymbol{v}}{\|\boldsymbol{v}\|} \tag{4.69}$$

which is in this case even species-dependent.

Using the Fickian law (4.67) and incorporating the liquid mass conservation (4.22), the species mass balance law (4.65) is written in its convective form

$$B\varepsilon s\frac{\partial C_k}{\partial t}+B\varepsilon s\boldsymbol{v}\cdot\nabla C_k-\nabla\cdot(B\varepsilon s\boldsymbol{D}_k\cdot\nabla C_k)+(\bar{Q}+B\varepsilon s\vartheta_k)C_k = B\varepsilon s\tilde{R}_k \tag{4.70}$$

Considering additionally sorption effects in the porous medium in accordance with Sect. 3.9.2 the following species mass transport equation can be derived:

$$B\varepsilon s\acute{\Re}_k\frac{\partial C_k}{\partial t} + B\varepsilon s\boldsymbol{v}\cdot\nabla C_k - \nabla\cdot(B\varepsilon s\boldsymbol{D}_k\cdot\nabla C_k) + (\bar{Q} + B\varepsilon s\vartheta_k\Re_k)C_k = B\varepsilon s\tilde{R}_k \tag{4.71}$$

with the retardation relationships

$$\begin{aligned}\Re_k &= 1 + \left(\tfrac{1-\varepsilon}{\varepsilon}\right)\varphi_k\\ \acute{\Re}_k &= 1 + \left(\tfrac{1-\varepsilon}{\varepsilon}\right)\tfrac{\partial(\varphi_k C_k)}{\partial C_k}\end{aligned} \tag{4.72}$$

in which the sorption function φ_k can be specified for Henry, Freundlich and Langmuir isotherms as listed in Table 3.8.

4.3.4 Energy Conservation

The energy balance equation is derived basically from (4.6) with Table 4.2 under the assumption of a thermal equilibrium between the liquid l and the solid s phase (see Sect. 3.9.2 for more details). We obtain finally (note that the liquid phase index is omitted for convenience):

$$\frac{\partial}{\partial t}\{B[\varepsilon s\rho E + (1-\varepsilon)\rho^s E^s]\} + \nabla\cdot(B\varepsilon s\rho E\boldsymbol{v}) + \nabla\cdot(B\boldsymbol{j}_T) = B\,H_e \tag{4.73}$$

with

$$H_e = \varepsilon s\rho H + (1-\varepsilon)\rho^s H^s \tag{4.74}$$

which can be applied to all interesting heat transport problems when specifying ε, s and B appropriately. Note that the generalized thermal source/sink term H_e includes both interfacial and surfacial heat transfer conditions.

Using the state relation (3.208) for the internal energy

$$dE^{\alpha} = c^{\alpha}\, dT \qquad \text{for} \quad (\alpha = l, s) \tag{4.75}$$

and the *Fourier heat flux* (3.176) with (3.177)

$$\begin{aligned} \boldsymbol{j}_T &= -\boldsymbol{\Lambda} \cdot \nabla T \\ \boldsymbol{\Lambda} &= \boldsymbol{\Lambda}_0 + \boldsymbol{\Lambda}_{\text{mech}} = [\varepsilon s \Lambda + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta} + \varepsilon s \rho c \boldsymbol{D}_{\text{mech}} \end{aligned} \tag{4.76}$$

it yields the following balance equation for the thermal energy written in its convective form:

$$\begin{aligned} B[\varepsilon s \rho c + (1-\varepsilon)\rho^s c^s]\frac{\partial T}{\partial t} + B\varepsilon s \rho c \boldsymbol{v} \cdot \nabla T - \nabla \cdot (B\, \boldsymbol{\Lambda} \cdot \nabla T) + \\ \rho c (T - T_0)\bar{Q} = B\, H_e \end{aligned} \tag{4.77}$$

to be solved for the system temperature T.

4.3.5 Generalized Model Equations

4.3.5.1 Flow

The fundamental flow equation represents a combination of the liquid mass conservation equation (4.27) and the liquid momentum conservations for porous media (4.38), Hagen-Poiseuille flow (4.51) and overland/channel flow (4.63). As the result, Table 4.5 summarizes the governing equation for 1D and 2D discrete features in dependence on the problem cases under consideration. For the Hagen-Poiseuille flow and overland/channel flow standard geometric forms of the fractures are embodied. Different geometries can be input by means of corrections in the corresponding hydraulic parameters as described in Sect. 4.4. We note that variably saturated conditions only exist for porous-medium flow while free surface (phreatic) conditions are only applicable to fully saturated porous media and pure liquid flow.

4.3.5.2 Species Mass

The governing species mass transport equation (4.71) can now be specified for the different flow conditions and discrete features. Table 4.6 summarizes the different terms and expressions for both porous media and pure liquid conditions.

4.3.5.3 Heat

The specified terms for the governing heat transport equation (4.77) are summarized in Table 4.7 for both porous media and pure liquid conditions.

4.4 Specifying Geometries and Hydraulic Radii for Different Types of Discrete Features

4.4.1 Standard Settings

Discrete features in 1D and 2D have cross-sectional geometries which are commonly input as flow area A and thickness/flow depth B, respectively (see Table 4.8). As outlined in Table 4.5 three laws of flow motion are provided for discrete features: Darcy flow in porous media as well as Hagen-Poiseuille and overland pure liquid flow. These laws require different input parameters of flow conductancy or friction, which have been summarized in Table 4.9. Note that for overland flow the Manning-Strickler law is preferred in the following.

It is obvious, the dataset of flow motion for the Hagen-Poiseuille and the Manning-Strickler laws needs in addition the prescription of the hydraulic radius r_{hydr}, which must be in accordance with the cross-sectional geometry of Table 4.8. To avoid redundancy and conflicts in the input dataset of discrete features it is preferred to use standard hydraulic radii r_{hydr} which are related to the dimension of the discrete features and the cross-sectional input geometry. The standard hydraulic radii r_{hydr} are summarized in Table 4.10. Those hydraulic radii r_{hydr} which are different from the standard ones can be specified via corrects in the frictional input parameters as described as follows.

4.4.2 Hydraulic Aperture of Hagen-Poiseuille Law

The standard hydraulic conductivity $\boldsymbol{K}$ of the Hagen-Poiseuille law is according to (4.51) and Table 4.10:

$$\boldsymbol{K} = \frac{r_{\text{hydr}}^2\,\rho_0 g}{a\,\mu_0}\,\boldsymbol{\delta} = \frac{b^2\,\rho_0 g}{12\mu_0}\,\boldsymbol{\delta} \tag{4.78}$$

where $a = 3$ for plane geometry. The following standard parameter set is used for the liquid of water: $\rho_0 = 10^3\,\text{kg}\,\text{m}^{-3}$, $\mu_0 = 1.3 \cdot 10^{-3}\,\text{Pa}\,\text{s}$ and $g = 9.81\,\text{m}\,\text{s}^{-2}$. It results a factor of $f_0 = \rho_0 g/\mu_0 = 7.55 \cdot 10^6\,\text{m}^{-1}\,\text{s}^{-1}$. A hydraulic radius, which is different from the standard geometry, and parameters, which are different from the standard parameter factor f_0, can be derived from the identity

Table 4.5 Flow model equations

$$\mathcal{L}(h(s)) = S\frac{\partial h}{\partial t} + O\frac{\partial s}{\partial t} - \nabla \cdot \big(B\boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})\big) - Q = 0$$

		S		O		$B\boldsymbol{K}$			
Dimension	Case	Porous	Pure liquid	Porous	Pure liquid	Darcy	Hagen-Poiseuille	Overland	Q
1D	PP[a]	$b(BS_o + \varepsilon_e)$	$b(B\rho_0 g\gamma + 1)$	0	0	$bB\underbrace{\frac{\boldsymbol{k}\rho_0 g}{\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$bB\underbrace{\frac{r_{\text{hydr}}^2\rho_0 g}{3\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$bB\underbrace{\frac{\tau r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\\|\nabla h\\|^2}}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$b\bar{Q}$
1D	PN[b]	bBS_o	$bB\rho_0 g\gamma$	$bB\varepsilon$	0	$bB\underbrace{\frac{\boldsymbol{k}\rho_0 g}{\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$bB\underbrace{\frac{r_{\text{hydr}}^2\rho_0 g}{3\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$bB\underbrace{\frac{\tau r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\\|\nabla h\\|^2}}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$b\bar{Q}$
1D	TP[c]	$\pi R^2(S_o + \frac{\varepsilon_e}{B})$	$\pi R^2(\rho_0 g\gamma + \frac{1}{B})$	0	0	$\pi R^2\underbrace{\frac{\boldsymbol{k}\rho_0 g}{\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\pi R^2\underbrace{\frac{r_{\text{hydr}}^2\rho_0 g}{2\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\pi R^2\underbrace{\frac{\tau r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\\|\nabla h\\|^2}}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\pi R^2\bar{Q}$
1D	TN[d]	$\pi R^2 S_o$	$\pi R^2\rho_0 g\gamma$	$\pi R^2\varepsilon$	0	$\pi R^2\underbrace{\frac{\boldsymbol{k}\rho_0 g}{\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\pi R^2\underbrace{\frac{r_{\text{hydr}}^2\rho_0 g}{2\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\pi R^2\underbrace{\frac{\tau r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\\|\nabla h\\|^2}}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\pi R^2\bar{Q}$
2D	PP[a]	$BS_o + \varepsilon_e$	$B\rho_0 g\gamma + 1$	0	0	$B\underbrace{\frac{\boldsymbol{k}\rho_0 g}{\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$B\underbrace{\frac{r_{\text{hydr}}^2\rho_0 g}{3\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$B\underbrace{\frac{\tau r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\\|\nabla h\\|^2}}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\bar{Q}$
2D	PN[b]	BS_o	$B\rho_0 g\gamma$	$B\varepsilon$	0	$B\underbrace{\frac{\boldsymbol{k}\rho_0 g}{\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$B\underbrace{\frac{r_{\text{hydr}}^2\rho_0 g}{3\mu_0}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$B\underbrace{\frac{\tau r_{\text{hydr}}^{\delta/2}}{\sqrt[4]{\\|\nabla h\\|^2}}\boldsymbol{\delta}}_{\boldsymbol{K}}$	$\bar{Q}$

[a] Plane phreatic
[b] Plane non-phreatic
[c] Tubular phreatic
[d] Tubular non-phreatic

Table 4.6 Species mass transport model equations

$$\mathcal{L}(C_k) = S_k \frac{\partial C_k}{\partial t} + \boldsymbol{q} \cdot \nabla C_k - \nabla \cdot (B \varepsilon s \boldsymbol{D}_k \cdot \nabla C_k) + \Phi_k C_k - Q_k = 0$$

		S_k		$\boldsymbol{q}$		$B\varepsilon s\boldsymbol{D}_k$		Φ_k		Q_k	
Dimension	Case	Porous	Pure liquid	Porous	Pure liquid	Porous	Pure liquid	Porous	Pure liquid	Porous	Pure liquid
1D	PP[a]	$bB\varepsilon\acute{\mathfrak{R}}_k$	bB	$bB\varepsilon\boldsymbol{v}$	$bB\boldsymbol{v}$	$bB\varepsilon(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$bB(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$b(\bar{Q} + B\varepsilon\vartheta_k\mathfrak{R}_k)$	$b(\bar{Q} + B\vartheta_k)$	$bB\varepsilon\tilde{R}_k$	$bB\tilde{R}_k$
1D	PN[b]	$bB\varepsilon s\acute{\mathfrak{R}}_k$	bB	$bB\varepsilon s\boldsymbol{v}$	$bB\boldsymbol{v}$	$bB\varepsilon s(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$bB(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$b(\bar{Q} + B\varepsilon s\vartheta_k\mathfrak{R}_k)$	$b(\bar{Q} + B\vartheta_k)$	$bB\varepsilon s\tilde{R}_k$	$bB\tilde{R}_k$
1D	TP[c]	$\pi R^2\varepsilon\acute{\mathfrak{R}}_k$	πR^2	$\pi R^2\varepsilon\boldsymbol{v}$	$\pi R^2\boldsymbol{v}$	$\pi R^2\varepsilon(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$\pi R^2(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$\pi R^2(Q + \varepsilon\vartheta_k\mathfrak{R}_k)$	$\pi R^2(Q + \vartheta_k)$	$\pi R^2\varepsilon\tilde{R}_k$	$\pi R^2\tilde{R}_k$
1D	TN[d]	$\pi R^2\varepsilon s\acute{\mathfrak{R}}_k$	πR^2	$\pi R^2\varepsilon s\boldsymbol{v}$	$\pi R^2\boldsymbol{v}$	$\pi R^2\varepsilon s(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$\pi R^2(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$\pi R^2(Q + \varepsilon s\vartheta_k\mathfrak{R}_k)$	$\pi R^2(Q + \vartheta_k)$	$\pi R^2\varepsilon s\tilde{R}_k$	$\pi R^2\tilde{R}_k$
2D	PP[a]	$B\varepsilon\acute{\mathfrak{R}}_k$	B	$B\varepsilon\boldsymbol{v}$	$B\boldsymbol{v}$	$B\varepsilon(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$B(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$\bar{Q} + B\varepsilon\vartheta_k\mathfrak{R}_k$	$\bar{Q} + B\vartheta_k$	$B\varepsilon\tilde{R}_k$	$B\tilde{R}_k$
2D	PN[b]	$B\varepsilon s\acute{\mathfrak{R}}_k$	B	$B\varepsilon s\boldsymbol{v}$	$B\boldsymbol{v}$	$B\varepsilon s(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$B(D_k\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}})$	$\bar{Q} + B\varepsilon s\vartheta_k\mathfrak{R}_k$	$\bar{Q} + B\vartheta_k$	$B\varepsilon s\tilde{R}_k$	$B\tilde{R}_k$

[a] Plane phreatic
[b] Plane non-phreatic
[c] Tubular phreatic
[d] Tubular non-phreatic

Table 4.7 Heat transport model equations

$$\mathcal{L}(T) = S\frac{\partial T}{\partial t} + \boldsymbol{q}\cdot\nabla T - \nabla\cdot(B\boldsymbol{\Lambda}\cdot\nabla T) + \Phi(T - T_0) - Q = 0$$

		S		$\boldsymbol{q}$		$B\boldsymbol{\Lambda}$		Φ		Q	
Dimension	Case	Porous	Pure liquid	Porous	Pure liquid	Porous	Pure liquid	Porous	Pure liquid	Porous	Pure liquid
1D	PP[a]	$bB[\varepsilon\rho c + (1-\varepsilon)\rho^s c^s]$	$bB\rho c$	$bB\varepsilon\rho c\boldsymbol{v}$	$bB\rho c\boldsymbol{v}$	$bB\{[\varepsilon\Lambda + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta} + \varepsilon\rho c\boldsymbol{D}_{\text{mech}}\}$	$bB(\Lambda\boldsymbol{\delta} + \rho c\boldsymbol{D}_{\text{mech}})$	$b\rho c\bar{Q}$	$b\rho c\bar{Q}$	$bB[\varepsilon\rho H + (1-\varepsilon)\rho^s H^s]$	$bB\rho H$
1D	PN[b]	$bB[\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s]$	$bB\rho c$	$bB\varepsilon s\rho c\boldsymbol{v}$	$bB\rho c\boldsymbol{v}$	$bB\{[\varepsilon s\Lambda + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta} + \varepsilon s\rho c\boldsymbol{D}_{\text{mech}}\}$	$bB(\Lambda\boldsymbol{\delta} + \rho c\boldsymbol{D}_{\text{mech}})$	$b\rho c\bar{Q}$	$b\rho c\bar{Q}$	$bB[\varepsilon s\rho H + (1-\varepsilon)\rho^s H^s]$	$bB\rho H$
1D	TP[c]	$\pi R^2[\varepsilon\rho c + (1-\varepsilon)\rho^s c^s]$	$\pi R^2\rho c$	$\pi R^2\varepsilon\rho c\boldsymbol{v}$	$\pi R^2\rho c\boldsymbol{v}$	$\pi R^2\{[\varepsilon\Lambda + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta} + \varepsilon\rho c\boldsymbol{D}_{\text{mech}}\}$	$\pi R^2(\Lambda\boldsymbol{\delta} + \rho c\boldsymbol{D}_{\text{mech}})$	$\pi R^2\rho c Q$	$\pi R^2\rho c Q$	$\pi R^2[\varepsilon\rho H + (1-\varepsilon)\rho^s H^s]$	$\pi R^2\rho H$
1D	TN[d]	$\pi R^2[\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s]$	$\pi R^2\rho c$	$\pi R^2\varepsilon s\rho c\boldsymbol{v}$	$\pi R^2\rho c\boldsymbol{v}$	$\pi R^2\{[\varepsilon s\Lambda + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta} + \varepsilon s\rho c\boldsymbol{D}_{\text{mech}}\}$	$\pi R^2(\Lambda\boldsymbol{\delta} + \rho c\boldsymbol{D}_{\text{mech}})$	$\pi R^2\rho c Q$	$\pi R^2\rho c Q$	$\pi R^2[\varepsilon s\rho H + (1-\varepsilon)\rho^s H^s]$	$\pi R^2\rho H$
2D	PP[a]	$B[\varepsilon\rho c + (1-\varepsilon)\rho^s c^s]$	$B\rho c$	$B\varepsilon\rho c\boldsymbol{v}$	$B\rho c\boldsymbol{v}$	$B\{[\varepsilon\Lambda + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta} + \varepsilon\rho c\boldsymbol{D}_{\text{mech}}\}$	$B(\Lambda\boldsymbol{\delta} + \rho c\boldsymbol{D}_{\text{mech}})$	$\rho c\bar{Q}$	$\rho c\bar{Q}$	$B[\varepsilon\rho H + (1-\varepsilon)\rho^s H^s]$	$B\rho H$
2D	PN[b]	$B[\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s]$	$B\rho c$	$B\varepsilon s\rho c\boldsymbol{v}$	$B\rho c\boldsymbol{v}$	$B\{[\varepsilon s\Lambda + (1-\varepsilon)\Lambda^s]\boldsymbol{\delta} + \varepsilon s\rho c\boldsymbol{D}_{\text{mech}}\}$	$B(\Lambda\boldsymbol{\delta} + \rho c\boldsymbol{D}_{\text{mech}})$	$\rho c\bar{Q}$	$\rho c\bar{Q}$	$B[\varepsilon s\rho H + (1-\varepsilon)\rho^s H^s]$	$B\rho H$

[a] Plane phreatic
[b] Plane non-phreatic
[c] Tubular phreatic
[d] Tubular non-phreatic

Table 4.8 Standard input of cross-sectional geometry for discrete features

Dimension	Fracture type/case	Parameter
1D	Plane, tubular (phreatic, non-phreatic)	Flow area, A
2D	Plane (phreatic, non-phreatic)	Thickness, B

Table 4.9 Frictional input parameters for discrete features

Law	Parameter
Darcy	Hydraulic conductivity, $\boldsymbol{K}$
Hagen-Poiseuille	Hydraulic aperture, b
Manning-Strickler	Roughness, M

Table 4.10 Standard hydraulic radii r_{hydr} used for discrete features

	r_{hydr}	
Dimension	Hagen-Poiseuille	Manning-Strickler
1D (plane)	$b/2$ (submerged slit plane, type (b) of Table 4.3)	$\sqrt{A}/4$ (submerged quadratic cross section)
2D (plane)		$B/2$ (submerged slit plane, type (b) of Table 4.3)

$$\frac{r_{\text{hydr}}^2}{a} f = \frac{b^2}{12} f_0 \quad \text{with} \quad f = \frac{\rho g}{\mu} \tag{4.79}$$

A *corrected hydraulic aperture* b_{corr} can be obtained from (4.79) as

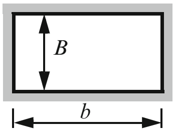
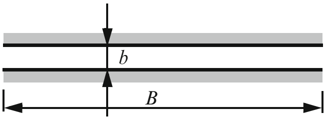
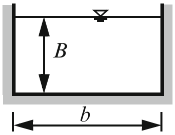
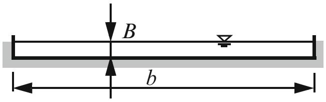
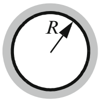
$$b_{\text{corr}} = \sqrt{\frac{f}{f_0} \frac{\sqrt{3}}{\sqrt{a}}}\, r_{\text{hydr}} \qquad a = \begin{cases} 3 \text{ for plane flow} \\ 2 \text{ for axisymmetric flow} \end{cases} \tag{4.80}$$

where r_{hydr} is the actual (true) hydraulic radius, which can be taken from Table 4.3, and f is the true parameter factor, where dynamic viscosity μ, gravity g and density ρ can be specified different from the standard settings in f_0. Table 4.11 summarizes the corrected apertures b_{corr} for interesting applications.

4.4.3 Roughness Coefficient of Manning-Strickler Law

The standard hydraulic conductivity $\boldsymbol{K}$ of the Manning-Strickler law is according to (4.63) with Tables 4.4 and 4.10:

Table 4.11 Corrected apertures b_{corr} for different applications in the case of Hagen-Poiseuille law for 1D and 2D discrete features[a]

Case	Type	b_{corr}	Remark
(a)	submerged rectangular cross - section  	$\frac{Bb}{(b+B)}\sqrt{\frac{f}{f_0}}$	
(b)	submerged slit plane  	$b\sqrt{\frac{f}{f_0}}$	see[b]
(c)	open rectangular cross - section  	$\frac{2Bb}{(b+2B)}\sqrt{\frac{f}{f_0}}$	
(d)	open wide channel (b>20 B), plane  	$2B\sqrt{\frac{f}{f_0}}$	
(e)	submerged circular cross - section  	$1.224745\,R\sqrt{\frac{f}{f_0}}$	

[a] $f_0 = \frac{\rho_0 g}{\mu_0} = 7.55 \cdot 10^6\,\mathrm{m}^{-1}\,\mathrm{s}^{-1}, \ f = \frac{\rho g}{\mu}$

[b] No correction is needed if $f = f_0$

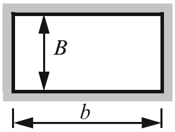
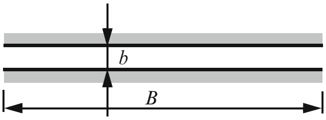
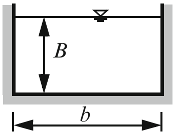
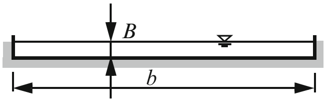
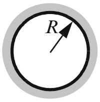
$$\boldsymbol{K} = M\, r_{\mathrm{hydr}}^{2/3} \frac{\boldsymbol{\delta}}{\sqrt[4]{\|\nabla h\|^2}} = \begin{cases} M\left(\frac{\sqrt{A}}{4}\right)^{2/3} \frac{\boldsymbol{\delta}}{\sqrt[4]{\|\nabla h\|^2}} & \text{for 1D} \\ M\left(\frac{B}{2}\right)^{2/3} \frac{\boldsymbol{\delta}}{\sqrt[4]{\|\nabla h\|^2}} & \text{for 2D} \end{cases} \tag{4.81}$$

Accordingly, from (4.81) we can find a *corrected Manning coefficient* M_{corr} in the following form to specify hydraulic radii r_{hydr}, which are different from the standard geometry of 1D and 2D discrete features:

$$M_{\mathrm{corr}} = M\, r_{\mathrm{hydr}}^{2/3} \begin{cases} \left(\frac{4}{\sqrt{A}}\right)^{2/3} & \text{for 1D} \\ \left(\frac{2}{B}\right)^{2/3} & \text{for 2D} \end{cases} \tag{4.82}$$

where M is the true (physical) Manning roughness coefficient. Table 4.12 summarizes the corrections M_{corr}/M for the Manning coefficients applied to 1D and 2D discrete features.

Table 4.12 Corrected Manning roughness coefficient M_{corr} for different applications in the case of Manning-Strickler law for 1D and 2D discrete features

		M_{corr}/M		
Case	Type	1D	2D	Remark
(a)	submerged rectangular cross - section 	$\left(\frac{2Bb}{(b+B)\sqrt{A}}\right)^{2/3}$	$\left(\frac{b}{b+B}\right)^{2/3}$	see[a]
(b)	submerged slit plane 	$\left(\frac{2b}{\sqrt{A}}\right)^{2/3}$	1	see[b]
(c)	open rectangular cross - section 	$\left(\frac{4Bb}{(b+2B)\sqrt{A}}\right)^{2/3}$	$\left(\frac{2b}{b+2B}\right)^{2/3}$	
(d)	open wide channel (b>20 B), plane 	$\left(\frac{4B}{\sqrt{A}}\right)^{2/3}$	$2^{2/3} = 1.5874$	
(e)	submerged circular cross - section 	$\left(\frac{2R}{\sqrt{A}}\right)^{2/3}$	-	

[a] Related to 1D: if $b = B$ and $A = B^2$ then $M_{\text{corr}} = M$, i.e., no correction is needed
[b] Related to 2D: it is $M_{\text{corr}} = M$, so no correction is required

Chapter 5
Chemical Reaction

5.1 General

The quantities r_k^α, R_k^α, R_k, $\tilde{R}_k$, $\bar{R}_k$ or $\bar{\tilde{R}}_k$ that appear in the species mass transport equations (3.50), (3.51), (3.248) and (4.71) and those of Tables 3.5, 3.7, 3.9–3.11 and 4.6 represent rates of production of mass of chemical species k due to chemical reactions occurring within a phase α, termed as *homogeneous reactions*, or between two or more phases, termed as *heterogeneous reactions*. Chemical reaction rate expressions have to be developed by constitutive relations as indicated in Sect. 3.8.5.6. Those expressions are most often determined from experimental studies and introduce rate parameters in form of constants and exponents, which are related to the concentrations of the chemical constituents involved in the reaction.

Chemical reactions are usually divided into *fast* and *slow* reactions. This type of classification is done in relation to the magnitude of the rate constant used in the chemical reaction rate expression. Thermodynamically, fast reactions are *reversible* and are locally in a thermodynamic equilibrium, while slow reactions represent *irreversible* reaction processes for which kinetic rate laws are required. Reversible and irreversible reactions have different meaning when referring to reaction kinetics and allow a rather different treatment in their mathematical description. An irreversible kinetic reaction is one which proceeds in only one direction, symbolized by $\rightarrow$, whereas a reversible kinetic reaction can proceed in the forward and backward directions, symbolized by $\rightleftharpoons$. Following types of reactions can be exemplified:

Binary ion exchange reaction between liquid (aqueous) and solid phase (adsorption isotherm)

$$A + B \rightleftharpoons C + D \tag{5.1}$$

First-order reaction (decay)

$$A \rightarrow P \tag{5.2}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_5,

Consecutive reaction (decay chains, serial reaction)

$$A \rightarrow B \rightarrow C \rightarrow D \tag{5.3}$$

Michaelis-Menten mechanism (Monod kinetics)

$$A + E \rightleftharpoons AE \rightarrow P + E \tag{5.4}$$

Parallel complex reactions

$$\begin{aligned} A + B + C &\rightarrow D + E \\ B + E + F &\rightarrow P \\ D + G &\rightleftharpoons H + I \\ I + A &\rightarrow J + K \end{aligned} \tag{5.5}$$

where $A, B, \ldots$ represent chemical species (reactants or products) which may be associated with different phases.

Whether a specific reaction is fast or slow depends on various physical properties of the reaction system. In general, a heterogeneous reaction process may consist of as many as five steps in series [386, 565]:

1. Diffusive transport of solute molecules to the interface.
2. Adsorption at the interface.
3. Reaction at the surface.
4. Desorption of products at the interface.
5. Diffusive transport of products from the interface.

If the steps 2–4 are faster than 1 and 5, the overall reaction is considered to be *transport controlled*. If the reverse is true, the overall reaction is said to be *surface controlled*.

5.2 Governing Mass Transport Equations

5.2.1 Balance Statements

The mass conservation of chemical species in the $\alpha-$phase can be concisely written in the following general balance equation encompassing all forms of our interest (cf. Chaps. 3 and 4):

$$\mathcal{L}_k^\alpha C_k^\alpha = \varepsilon_\alpha (r_k^\alpha + R_k^\alpha) \tag{5.6}$$

with

$$\mathcal{L}_k^\alpha\, C_k^\alpha = \begin{cases} \dfrac{\partial}{\partial t}(\varepsilon_\alpha C_k^\alpha) + \nabla \cdot (\boldsymbol{q}^\alpha C_k^\alpha) + \nabla \cdot \boldsymbol{j}_{\alpha k} & \begin{cases} \alpha = \text{liquid } l \\ k = \text{liquid species} \end{cases} \\ \dfrac{\partial}{\partial t}(\varepsilon_\alpha C_k^\alpha) & \begin{cases} \alpha = \text{solid } s \\ k = \text{solid species} \end{cases} \end{cases} \tag{5.7}$$

where each species, labeled by the subscript k, is associated with a particular phase $\alpha \in (l, s)$, where l and s indicate the liquid and the solid phase, respectively. By definition, $\boldsymbol{q}^\alpha = \boldsymbol{j}_{\alpha k} = \boldsymbol{0}$ for the solid phase s. In (5.6) and (5.7), $\mathcal{L}_k^\alpha$ is a differential operator, C_k^α is the concentration of species k of $\alpha-$phase (3.266), ε_α is the volume fraction of $\alpha-$phase (3.4), r_k^α and R_k^α are the homogeneous and heterogeneous reaction rates of species k of $\alpha-$phase, respectively, $\boldsymbol{q}^\alpha$ is the Darcy velocity of $\alpha-$phase (3.240) and $\boldsymbol{j}_{\alpha k} = \varepsilon_\alpha \boldsymbol{j}_k^\alpha$ is the bulk mass flux ($\boldsymbol{j}_k^\alpha$ is the intrinsic mass flux) of species k of $\alpha-$phase (3.272).

By using (3.219) the volume fraction ε_α can be expressed for the liquid l and solid s phases, respectively, as

$$\begin{aligned} \varepsilon_l &= \varepsilon\, s^l \\ \varepsilon_s &= 1 - \varepsilon \end{aligned} \tag{5.8}$$

where ε is the porosity (void space) and s^l is the saturation referring to the liquid phase l. For unsaturated porous media it is $s^l < 1$, whereas for saturated porous media we have $s^l = 1$ and $\varepsilon_l = \varepsilon$. In case of a pure liquid it is simply $\varepsilon_l = 1$. With (5.8) the balance equation (5.6) can be written for the liquid and solid phases, respectively,

$$\begin{aligned} \mathcal{L}_k^l\, C_k^l &= \frac{\partial}{\partial t}(\varepsilon s^l C_k^l) + \nabla \cdot (\boldsymbol{q}^l C_k^l) + \nabla \cdot \boldsymbol{j}_{lk} = \varepsilon s^l (r_k^l + R_k^l) \\ \mathcal{L}_k^s\, C_k^s &= \frac{\partial}{\partial t}(\varepsilon_s C_k^s) = \varepsilon_s (r_k^s + R_k^s) \end{aligned} \tag{5.9}$$

5.2.2 Reaction Rates and Multiphase Aspects

The solution of the balance equation (5.6) requires knowledge of the reaction rates for kinematically controlled reactions. Different forms of rate laws can be derived (cf. Sect. 3.8.5.6). These forms depend on the type of reaction and whether the reaction is homogeneous or heterogeneous. In general, if a species k exists in more than one phase α, for instance the species is exchanged between liquid l and solid s phases in an adsorption process, the transport equation (5.6) has to be summed over all contributed phases α:

$$\sum_\alpha (\mathcal{L}_k^\alpha\, C_k^\alpha) = R_k \tag{5.10}$$

with the bulk reaction rate (3.189)

$$R_k = R_{\text{hom}_k} + R_{\text{het}_k} = \underbrace{\sum_\alpha \varepsilon_\alpha r_k^\alpha}_{\text{homogeneous}} + \underbrace{\sum_\alpha \varepsilon_\alpha R_k^\alpha}_{\text{heterogeneous}} \tag{5.11}$$

where $R_{\text{hom}_k} = \sum_\alpha \varepsilon_\alpha r_k^\alpha$ is the homogeneous reaction rate and $R_{\text{het}_k} = \sum_\alpha \varepsilon_\alpha R_k^\alpha$ is the heterogeneous reaction rate of species k accumulating over all its contributing phases α.

In contrast, if a species k exists only in one phase, say phase α, then (5.10) is reduced to

$$\mathcal{L}_k^\alpha C_k^\alpha = R_k \tag{5.12}$$

with

$$R_k = R_{\text{hom}_k} + R_{\text{het}_k} = \underbrace{\varepsilon_\alpha r_k^\alpha}_{\text{homogeneous}} + \underbrace{\varepsilon_\alpha R_k^\alpha}_{\text{heterogeneous}} \tag{5.13}$$

Because the sum of the reacting mass must be identical to the sum of the produced mass in all cases, the reaction rates are governed by the constraint:

$$\sum_i (R_{\text{hom}_k} + R_{\text{het}_k}) = 0 \tag{5.14}$$

The transport processes of interest refer to a liquid phase l (solvent) moving through a porous medium in which the void space is variably saturated by the $l-$phase. Conceptually, a variably saturated medium consists, at least, of three phases: liquid l, gas g and solid s (see Sect. 3.8.7). Thus, we have for the volume fractions (3.219) and (3.220)

$$1 = \underbrace{\varepsilon_l + \varepsilon_g}_{\varepsilon} + \varepsilon_s \tag{5.15}$$

Since the $l-$phase occupies only part of the porosity (void space) ε, the saturation s^l of the $l-$phase (3.219) defines the relative quantity as

$$s^l = \frac{\varepsilon_l}{\varepsilon} \qquad (0 < s^l \leq 1) \tag{5.16}$$

Under *unsaturated conditions* $s^l \leq 1$ the liquid l as the wetting phase can occupy only part of the void space ε and therefore only part of the total area of the solid s can be exposed to exchange reactions in an adsorption process. Accordingly, we can subdivide the solid volume fraction ε_s into chemically active and inactive parts of solid mass, viz.,

$$\varepsilon_s = \varepsilon_{s^{\text{active}}} + \varepsilon_{s^{\text{inactive}}} \tag{5.17}$$

Obviously, the portion of the total surface of the solid that is in contact with the $l-$phase depends on ε_l. It can be assumed [39] 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^l. Thus,

$$\frac{A_{s^{\text{active}}}}{A_{s^{\text{active}}} + A_{s^{\text{inactive}}}} = \frac{\varepsilon_{s^{\text{active}}}}{\varepsilon_s} = f(\varepsilon_l) \approx \frac{\varepsilon_l}{\varepsilon} = s^l \tag{5.18}$$

and we obtain

$$\varepsilon_{s^{\text{active}}} = f(\varepsilon_l)\varepsilon_s \approx s^l \varepsilon_s = s^l(1-\varepsilon) \tag{5.19}$$

In contrast to this, it may be argued [38, 422] that l as the wetting phase completely coats the solid such that the nonwetting gas phase g is not in contact with the solid. Under these conditions it is assumed that the complete surface of the solid is exposed to exchange reactions, i.e., $\varepsilon_s = \varepsilon_{s^{\text{active}}}$. To take into account these different assumptions regarding the possible liquid-solid contact areas exposed to exchange reactions, we can write for the solid volume fraction

$$\varepsilon_s = \varepsilon_{s^{\text{active}}} = [s^l + \omega(1-s^l)](1-\varepsilon) \qquad \omega \in (0,1) \tag{5.20}$$

where ω introduces a *coating factor*, which is unity if a full exchange contact is assumed, otherwise if $\omega = 0$ the exchange contact relation amounts to the liquid saturation s^l. For the most practical applications we prefer the latter case with $\omega = 0$ and use

$$\varepsilon_s = \varepsilon_{s^{\text{active}}} = s^l(1-\varepsilon) \tag{5.21}$$

as the chemically active exchange fraction in the summed mass balance (5.10) written for the liquid and solid phases

$$\begin{aligned} \mathcal{L}_k^l C_k^l + \mathcal{L}_k^s C_k^s &= \underbrace{s^l[\varepsilon(r_k^l + R_k^l) + (1-\varepsilon)(r_k^s + R_k^s)]}_{R_k} \\ \frac{\partial}{\partial t}[\varepsilon s^l C_k^l + s^l(1-\varepsilon)C_k^s] + \nabla\cdot(\boldsymbol{q}^l C_k^l) + \nabla\cdot\boldsymbol{j}_{lk} &= R_k \end{aligned} \tag{5.22}$$

For given number of chemical reactions, $r = 1, \ldots, N_r$, the bulk reaction rate R_k for a species k, (5.11) or (5.13), can be expressed in the following general form [348]

$$R_k = \sum_{r=1}^{N_r} \nu_{kr} r_r \qquad (k = 1, \ldots, N) \tag{5.23}$$

where ν_{kr} is the *stoichiometric coefficient* of species k and reaction r and r_r is the bulk rate of reaction associated with the type of reaction r.

5.3 Basic Chemical Kinetics

5.3.1 Reaction Stoichiometry

The basis of chemical modeling represents the equations of reactions r, which can be written in their general stoichiometric form:
Forward reactions

$$\begin{aligned} |\nu_{1r}|A_1 + |\nu_{2r}|A_2 + \ldots + |\nu_{N^o r}|A_{N^o} \overset{k^+}{\rightarrow} |\nu_{(N^o+1)r}|B_{(N^o+1)} +& \\ |\nu_{(N^o+2)r}|B_{(N^o+2)} + \ldots + |\nu_{Nr}|B_N& \end{aligned} \tag{5.24}$$

Backward reactions

$$\begin{aligned} |\nu_{1r}|A_1 + |\nu_{2r}|A_2 + \ldots + |\nu_{N^o r}|A_{N^o} \overset{k^-}{\leftarrow} |\nu_{(N^o+1)r}|B_{(N^o+1)} +& \\ |\nu_{(N^o+2)r}|B_{(N^o+2)} + \ldots + |\nu_{Nr}|B_N& \end{aligned} \tag{5.25}$$

for $N^o < N$ $(N^o = \text{number of reactants})$ and $r = 1, \ldots, N_r$. They are related and quantified by the stoichiometric coefficients $|\nu_{kr}|$. The algebraic stoichiometric numbers ν_{kr} behave:

$$\begin{array}{llll} \nu_{kr} < 0 & \text{for} & 1 \leq k \leq N^o & \text{(reactants)} \\ \nu_{kr} > 0 & \text{for} & N^o \leq k \leq N & \text{(products)} \end{array} \tag{5.26}$$

In (5.24) and (5.25) A_k and B_k represent chemical species of reactants and products, respectively, and k^+ and k^- indicate rate constants for the forward and backward reactions, respectively. We note that the species A_k and B_k are generally associated with a phase, in particular the liquid phase l or the solid phase s. To emphasize their phase-relations we shall sometimes denote the species by the phase index in form of A_k^l, A_k^s or $A_k(aq)$, $A_k(s)$, referring to the liquid (aqueous) and solid phases, respectively. Furthermore, we denote as *monovalence* if $|\nu_{kr}| = 1$ for all species k in a reaction r, otherwise we denote as *heterovalence* if $|\nu_{kr}| > 1$ at least for one species k in a reaction r.

5.3.2 Examples

To illustrate the symbolic reaction stoichiometry (5.24) and (5.25) we exemplify the following irreversible and reversible reactions occurring in different applications of porous-media mass transport.

(i) *Pyrite oxidation*
One mechanism involves oxidation of pyrite by O_2. Another possible mechanism for the oxidation of pyrite is the reaction with Fe(III) as the oxidant. These irreversible reactions for the pyrite oxidation read [500]:

$$\begin{aligned} FeS_2 + \tfrac{7}{2}O_2 + H_2O &\to Fe^{2+} + 2SO_4^{2-} + 2H^+ \\ FeS_2 + 14Fe^{3+} + 8H_2O &\to 15Fe^{2+} + 2SO_4^{2-} + 16H^+ \end{aligned} \tag{5.27}$$

(ii) *Equilibrium reactions of carbonate system*
The chemical system contains $N = 7$ species:

$H_2O, OH^-, H^+, H_2CO_3, HCO_3^-, CO_3^{2-}, CO_2(aq)$

which are subjected to the following $N_r = 4$ equilibrium reactions [38]:

$$\begin{aligned} H_2O &\rightleftharpoons H^+ + OH^- \\ H_2CO_3 &\rightleftharpoons HCO_3^- + H^+ \\ HCO_3^- &\rightleftharpoons CO_3^{2-} + H^+ \\ CO_2(aq) + H_2O &\rightleftharpoons H_2CO_3 \end{aligned} \tag{5.28}$$

(iii) *Aerobic biodegeneration of BTEX*
The overall aerobic reaction stoichiometry for a fuel hydrocarbon (e.g., benzene) can be written as [86, 87]

$$C_6H_6 + 7.5O_2 \to 6CO_2 + 3H_2O \tag{5.29}$$

(iv) *Degradation of BTEX using multiple electron acceptors*
The biodegradation of BTEX can occur via five different degradation pathways [86, 442]: aerobic respiration, denitrification, iron reduction, sulfate reduction and methanogenesis. Accordingly, the following instantaneous five irreversible reactions are given ($N_r = 5$):

$$\begin{aligned} C_6H_6 + 7.5O_2 &\to 6CO_2 + 3H_2O \\ 6NO_3^- + 6H^+ + C_6H_6 &\to 6CO_2 + 6H_2O + 3N_2 \\ 30Fe(OH)_3 + 60H^+ + C_6H_6 &\to 6CO_2 + 78H_2O + 30Fe^{2+} \\ 3.75SO_4^{2-} + 7.5H^+ + C_6H_6 &\to 6CO_2 + 3H_2O + 3.75H_2S \\ C_6H_6 + 4.5H_2O &\to 2.25CO_2 + 3.75CH_4 \end{aligned} \tag{5.30}$$

(v) *Leaching of low-grade uranium ores*
Two principal types of low-grade uranium ores are uraninite (UO_2) and pitchblende (U_3O_8). Typical reaction equations may be written as [386]:

$$\begin{aligned} UO_2(s) + \tfrac{1}{2}O_2 + CO_3^{2-} + 2HCO_3^- &\rightarrow UO_2(CO_3)_3^{4-} + H_2O \\ U_3O_8(s) + \tfrac{1}{2}O_2 + 3CO_3^{2-} + 6HCO_3^- &\rightarrow 3UO_2(CO_3)_3^{4-} + 3H_2O \end{aligned} \qquad (5.31)$$

consisting of four reactants ($N^o = 4$) for each reaction.

(vi) *Radionuclide decay chain of uranium*
The radionuclide decay of ^{238}U occurs in the following decay series of serial and parallel reactions (note that U – uranium, Th – thorium, Pa – protactinium, Ra – radium, Rn – radon, Po – polonium, Pb – lead, Bi – bismuth, At – astatine, Tl – thallium) [183]:

$$\begin{tikzcd}
 & \underset{4.47\cdot 10^{9}\text{y}}{{}^{238}\text{U}} \arrow[r] & \underset{24.1\text{d}}{{}^{234}\text{Th}} \arrow[r] & \underset{1.17\text{m}}{{}^{234}\text{Pa}^{m}} \arrow[r] & \underset{6.75\text{h}}{{}^{234}\text{Pa}} \arrow[r] & \underset{2.45\cdot 10^{5}\text{y}}{{}^{234}\text{U}} \arrow[r] & \underset{7.7\cdot 10^{4}\text{y}}{{}^{230}\text{Th}} \arrow[r] & \underset{1600\text{y}}{{}^{226}\text{Ra}} \arrow[r] & {} \\
{} \arrow[r] & \underset{3.823\text{d}}{{}^{222}\text{Rn}} \arrow[r] & \underset{3.05\text{m}}{{}^{218}\text{Po}} \arrow[r, "0.02\%"] \arrow[d, "99.98\%"'] & \underset{2\text{s}}{{}^{218}\text{At}} \arrow[d] & & & & & \\
 & & \underset{26.8\text{m}}{{}^{214}\text{Pb}} \arrow[r] & \underset{19.7\text{m}}{{}^{214}\text{Bi}} \arrow[r] \arrow[d] & \underset{164\mu\text{s}}{{}^{214}\text{Po}} \arrow[d] & & & & \\
 & & & \underset{1.3\text{m}}{{}^{210}\text{Tl}} \arrow[r] & \underset{21\text{y}}{{}^{210}\text{Pb}} \arrow[r] & \underset{5.01\text{d}}{{}^{210}\text{Bi}} \arrow[r] & \underset{138.4\text{d}}{{}^{210}\text{Po}} \arrow[r] & \underset{\text{stable}}{{}^{206}\text{Pb}} &
\end{tikzcd} \qquad (5.32)$$

5.3.3 Rate Laws and Rate Constants

Based on the stoichiometric forms (5.24) and (5.25) the bulk reaction rates r_r in (5.23) for the forward ($r = 1$) and backward ($r = 2$) reactions can be expressed by the rate laws

$$\begin{aligned} r_1^+ &= k^+ \prod_{k=1}^{N^o} \varepsilon_{\alpha_k} \{A_k^\alpha\}^{|\nu_{k1}|} \\ r_2^- &= k^- \prod_{k=N^o+1}^{N} \varepsilon_{\alpha_k} \{B_k^\alpha\}^{|\nu_{k2}|} \end{aligned} \qquad \alpha \in (l, s) \qquad (5.33)$$

where k^+ and k^- represent rate constants. To emphasize which phase α contains the species k we use the specific phase index α_k, i.e., $\alpha = l$ if species k is dissolved in the liquid phase and $\alpha = s$ if species k is sorbed in the solid phase. The curly bracket symbol $\{\}$ refers to the (chemical) activity of the kth species at the $\alpha-$phase defined by (2.119).

5.3.4 Chemical Equilibrium and Law of Mass Action (LMA)

Chemical equilibrium describes a situation in which forward and backward reactions (5.24) and (5.25), respectively, are equal. It means

$$R_k = \sum_{r=1}^{N_r} \nu_{kr} r_r = 0 \qquad \forall\, k \tag{5.34}$$

Since

$$\begin{aligned} \nu_{k1} &= -|\nu_{k1}| && \text{for reactants} \\ \nu_{k2} &= +|\nu_{k2}| && \text{for products} \end{aligned} \tag{5.35}$$

it gives with (5.33)

$$K_{\text{eq}} = \frac{|\nu_{k1}|k^+}{|\nu_{k2}|k^-} = \frac{\prod\limits_{k=N^o+1}^{N} \varepsilon_{\alpha_k}\{B_k^\alpha\}^{|\nu_{k2}|}}{\prod\limits_{k=1}^{N^o} \varepsilon_{\alpha_k}\{A_k^\alpha\}^{|\nu_{k1}|}} \qquad \alpha \in (l, s) \tag{5.36}$$

Expression (5.36) is known as the *law of mass action* (LMA), where K_{eq} is the *equilibrium constant*. The equilibrium constant K_{eq} is to be known and measurable for given equilibrium reactions 1 and 2. For example, considering the simple binary ion monovalent exchange reaction in the form of (5.1) written as

$$A^l + B^s \underset{k^-}{\overset{k^+}{\rightleftarrows}} C^s + D^l \tag{5.37}$$

the LMA expression (5.36) yields

$$K_{\text{eq}} = \frac{k^+}{k^-} = \frac{\{C^s\}\{D^l\}}{\{A^l\}\{B^s\}} \tag{5.38}$$

5.3.5 Steady-State Approximation (SSA)

The steady-state approximation (SSA) [13] can be used to simplify the reaction analysis. It is supposed that the reaction rate of an intermediate species ($k = \text{int}$) is negligibly small, so that

$$R_{\text{int}} \approx 0 \tag{5.39}$$

For example, considering the consecutive monovalent reaction (5.3) of the form

$$A^l \overset{k_A}{\to} B^l \overset{k_B}{\to} C^l \tag{5.40}$$

where k_A and k_B are (forward) reaction constants, the SSA applied to species B^l becomes (with (5.23) and (5.33))

$$R_B = \varepsilon_l \big(k_A\{A^l\} - k_B\{B^l\}\big) \approx 0 \tag{5.41}$$

Then

$$\{B^l\} = \frac{k_A}{k_B}\{A^l\} \tag{5.42}$$

which can be used to express the reaction rate for species C^l

$$R_C = \varepsilon_l k_B\{B^l\} = \varepsilon_l k_A\{A^l\} \tag{5.43}$$

5.3.6 Pre-equilibria

Considering the following consecutive reaction

$$A + B \underset{k_A^-}{\overset{k_A^+}{\rightleftarrows}} C \overset{k_B}{\to} P \tag{5.44}$$

where C represents an intermediate species. This scheme involves a *pre-equilibrium* when the rates of formation of the intermediate C and its decay back into reactants A and B are much faster than its rate of formation of products P. Applying the SSA to species C of (5.44) it yields (exemplified for a homogeneous reaction in the l−phase)

$$R_C = \varepsilon_l \big(k_A^+\{A\}\{B\} - k_A^-\{C\} - k_B\{C\}\big) \approx 0 \tag{5.45}$$

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 (5.45) vanishes and we obtain

$$K_{\text{eq}} = \frac{k_A^+}{k_A^-} = \frac{\{C\}}{\{A\}\{B\}} \tag{5.46}$$

Then, the reaction rate for species C takes the form:

$$R_C = -\varepsilon_l k_B\{C\} = -\varepsilon_l k_B K_{\text{eq}}\{A\}\{B\} \tag{5.47}$$

which represents a second-order reaction law.

In extension to (5.44), let us consider the following reaction system

$$\begin{aligned} A + B &\underset{k_A^-}{\overset{k_A^+}{\rightleftarrows}} C + D \\ A &\overset{k_B}{\to} P \end{aligned} \tag{5.48}$$

It can be simplified with the pre-equilibrium assumption, where for instance the reaction rate of A results

$$R_A = -\varepsilon_l \underbrace{\left(k_A^+\{A\}\{B\} - k_A^-\{C\}\{D\}\right)}_{\Rightarrow K_{\text{eq}}=\frac{k_A^+}{k_A^-}=\frac{\{C\}\{D\}}{\{A\}\{B\}}} - \varepsilon_l k_B \{A\} \approx -\varepsilon_l k_B \{A\} \tag{5.49}$$

5.4 Selected Reaction Processes

5.4.1 Binary Exchange Reactions (Adsorption Isotherms)

Consider the heterogeneous reversible binary ion exchange reaction between the dissolved species A_1^l of the liquid phase l and the sorbed species A_2^s of the solid phase s in the form

$$|\nu_1| A_1^l + |\nu_2| A_2^s \rightleftharpoons |\nu_1| A_1^s + |\nu_2| A_2^l \tag{5.50}$$

where ν_1 and ν_2 are stoichiometric coefficients assumed to be independent of the direction of the reaction. At the equilibrium the LMA (5.36) yields

$$K_{\text{eq}} = \frac{\{A_1^s\}^{|\nu_1|}\{A_2^l\}^{|\nu_2|}}{\{A_1^l\}^{|\nu_1|}\{A_2^s\}^{|\nu_2|}} = \frac{(\gamma_1[A_1^s])^{|\nu_1|}(\gamma_2[A_2^l])^{|\nu_2|}}{(\gamma_1[A_1^l])^{|\nu_1|}(\gamma_2[A_2^s])^{|\nu_2|}} \tag{5.51}$$

where the activities { } are replaced by the molar concentrations [ ] of the species by using (2.119). Introducing the *ion exchange capacity* $[A_T^s]$ for the sorbed species in the form

$$[A_T^s] = \sum_k [A_k^s] = [A_1^s] + [A_2^s] \tag{5.52}$$

and the *total solution normality* $[A_T^l]$ for the liquid phase l as

$$[A_T^l] = \sum_k [A_k^l] = [A_1^l] + [A_2^l] \tag{5.53}$$

assuming, e.g., [549], that

- Dilute solutions occur so that $\gamma_1 = \gamma_2 = 1$,
- The ion exchange capacity $[A_T^s]$ is constant, and
- The total solution normality $[A_T^l]$ is also constant,

the following exchange relationships in form of adsorption isotherms can be derived from the equilibrium (5.51).

5.4.1.1 Langmuir Adsorption Isotherm

For the case of monovalence with $|\nu_1| = |\nu_2| = 1$, the condition of equilibrium (5.51) can be used to explicitly express the concentration of the sorbed solid species $[A_1^s]$ as a function of concentration of the dissolved species $[A_1^l]$, where the concentration of species $[A_2^s]$ and $[A_2^l]$ are substituted by inserting the ion exchange capacity $[A_T^s]$ and total solution normality $[A_T^l]$, respectively. After some manipulations we find

$$[A_1^s] = K_{\text{eq}} \frac{[A_T^s]}{[A_T^l]} \frac{[A_1^l]}{1 + \left(\dfrac{K_{\text{eq}} - 1}{[A_T^l]}\right)[A_1^l]} \tag{5.54}$$

Expressing the molar concentrations $[A_1^s]$ and $[A_1^l]$ by mass concentrations C_1^s and C_1^l, respectively, via (2.118), the following relationship results from (5.54)

$$C_1^s = \frac{k_1^\dagger C_1^l}{1 + k_1^\ddagger C_1^l} \tag{5.55}$$

which is termed the *Langmuir adsorption isotherm*, where $k_1^\dagger$ and $k_1^\ddagger$ are sorption coefficients defined by

$$k_1^\dagger = K_{\text{eq}} \frac{[A_T^s]}{[A_T^l]} \qquad k_1^\ddagger = \frac{K_{\text{eq}} - 1}{[A_T^l] m_1} \tag{5.56}$$

where m_1 is the molecular mass of species A_1^l.

5.4.1.2 Henry Adsorption Isotherm

Admitting for low concentration $k_l^\ddagger C_1^l \ll 1$ in the Langmuir isotherm (5.55) the well-known *Henry adsorption isotherm* results

$$C_1^s = \kappa_1 C_1^l \qquad (\kappa_1 = k_1^\dagger) \tag{5.57}$$

which provides a simple linear relation between the sorbed and the dissolved species for a monovalent binary exchange reaction, where κ_1 is the Henry sorption coefficient of species '1'. We note, $\kappa_1 = k_1^\dagger = K_{\text{eq}}$ when $[A_T^s]/[A_T^l] \approx 1$. The Henry coefficient κ_1 is often expressed by the *distribution coefficient* K_1^d (related to species '1')

$$K_1^d = \kappa_1 / \rho_s \tag{5.58}$$

where ρ_s is the density of the solid, so that

$$C_1^s = K_1^d \rho_s \, C_1^l \tag{5.59}$$

5.4.1.3 Freundlich Adsorption Isotherm

In case of heterovalent equilibrium reaction with $|\nu_1| = n \geq 1$ and $|\nu_2| = m \geq 1$ the equilibrium expression (5.51) is

$$K_{\text{eq}} = \left(\frac{[A_1^s]}{[A_1^l]} \right)^n \left(\frac{[A_2^l]}{[A_2^s]} \right)^m \tag{5.60}$$

and gets a polynomial relationship with respect to the sorbed solid species $[A_1^s]$ in the form

$$[A_1^s] = K_{\text{eq}}^{1/n} \left(\frac{[A_T^s] - [A_1^s]}{[A_T^l] - [A_1^l]} \right)^{m/n} [A_1^l] \tag{5.61}$$

However, explicit expressions are only possible for special cases. For example, having $n = 2$ and $m = 1$ we obtain from (5.61)

$$[A_1^s] = \frac{K_{\text{eq}}}{2([A_T^l] - [A_1^l])} \left(\sqrt{1 + \frac{4[A_T^s]([A_T^l] - [A_1^l])}{K_{\text{eq}}[A_1^l]^2}} - 1 \right) [A_1^l]^2 \tag{5.62}$$

providing a polynomial character $[A_1^s] \sim [A_1^l]^n$ of the isotherm. Often, however, simple empirical functions are preferred for this type of heterovalent equilibrium reaction to prevent the complexity in the direct evaluation of (5.60) for general cases. Here, most common is the *Freundlich adsorption isotherm* which simply reads

$$C_1^s = b_1^\dagger (C_1^l)^{b_1^\ddagger} \tag{5.63}$$

written for mass concentrations, where $b_1^\dagger$ is the Freundlich coefficient and $b_1^\ddagger \geq 1$ is the Freundlich exponent. Note that for $b_1^\ddagger \equiv 1$ the Freundlich isotherm becomes a Henry isotherm (5.57).

5.4.1.4 Adsorption Function and Retardation

To generalize the formulations of the adsorption isotherms of Sects. 5.4.1.1–5.4.1.3 we introduce the dimensionless *adsorption function* $\varphi_k = \varphi_k(C_k^l)$ for a species k to relate the solid species by the dissolved species for the Henry isotherm (5.57), the Freundlich isotherm (5.63) and the Langmuir isotherm (5.55) in the following form

$$C_k^s = \varphi_k \; C_k^l \tag{5.64}$$

with

$$\varphi_k = \begin{cases} \kappa_k & \text{Henry} \\ b_k^{\dagger} \, (C_k^l)^{b_k^{\ddagger}-1} & \text{Freundlich} \\ \frac{k_k^{\dagger}}{1+k_k^{\ddagger} C_k^l} & \text{Langmuir} \end{cases} \tag{5.65}$$

Using the mass balance equation (5.22) for species k summed over the liquid l and solid s phases

$$\frac{\partial}{\partial t}[\varepsilon s^l C_k^l + s^l (1-\varepsilon) C_k^s] + \nabla \cdot (\boldsymbol{q}^l C_k^l) + \nabla \cdot \boldsymbol{j}_{lk} = R_k \tag{5.66}$$

written in its divergence form and

$$\varepsilon s^l \frac{\partial C_k^l}{\partial t} + s^l (1-\varepsilon) \frac{\partial C_k^s}{\partial t} + \boldsymbol{q}^l \cdot \nabla C_k^l + \nabla \cdot \boldsymbol{j}_{lk} = R_k - C_k^l Q_l \tag{5.67}$$

written in its convective form (see also Sect. 3.9.2), we can replace the solid concentration C_k^s by the liquid concentration C_k^l via the adsorption relation (5.64) to obtain

$$\frac{\partial}{\partial t}\big(\varepsilon s^l \Re_k C_k^l\big) + \nabla \cdot (\boldsymbol{q}^l C_k^l) + \nabla \cdot \boldsymbol{j}_{lk} = R_k \tag{5.68}$$

written in the divergence form and

$$\varepsilon s^l \acute{\Re}_k \frac{\partial C_k^l}{\partial t} + \boldsymbol{q}^l \cdot \nabla C_k^l + \nabla \cdot \boldsymbol{j}_{lk} = R_k - C_k^l Q_l \tag{5.69}$$

written in the convective form, where

$$\Re_k = 1 + \Big(\frac{1-\varepsilon}{\varepsilon}\Big) \varphi_k \tag{5.70}$$

is the *retardation factor* and

$$\acute{\Re}_k = 1 + \left(\frac{1-\varepsilon}{\varepsilon}\right)\frac{\partial(\varphi_k C_k^l)}{\partial C_k^l} \tag{5.71}$$

is the *derivative term of retardation*, which are listed in Table 3.8 for the Henry, Freundlich and Langmuir isotherms.

5.4.2 First-Order Decay Reactions

Additional to the reversible heterogeneous exchange reaction (5.50), the species k in the liquid phase l and the solid phase s should be subjected to an irreversible homogeneous first-order decay reaction into the products P_k^l and P_k^s according to

$$\begin{aligned} |\nu_1|A_k^l + |\nu_2|A_m^s &\rightleftharpoons |\nu_1|A_k^s + |\nu_2|A_m^l \\ A_k^l &\overset{k_A}{\rightarrow} P_k^l \\ A_k^s &\overset{k_A}{\rightarrow} P_k^s \end{aligned} \tag{5.72}$$

where $m \neq k$ is a different species. We assume that the reversible heterogeneous reaction is much faster than the decay reactions. Under such conditions the pre-equilibrium assumption (5.49) becomes applicable. Applying (5.23), (5.33), (5.64) and (5.70), it leads to the following reaction rate

$$\begin{aligned} R_k &= -\varepsilon_l k_A\{A_k^l\} - \varepsilon_{s^{\text{active}}} k_A\{A_k^s\} \\ &= -s^l \underbrace{\frac{k_A\gamma_k}{m_k}}_{\vartheta_k}\left[\varepsilon C_k^l + (1-\varepsilon)C_k^s\right] \\ &= -s^l\vartheta_k\left[\varepsilon + (1-\varepsilon)\varphi_k\right]C_k^l \\ &= -\varepsilon s^l\vartheta_k\Re_k C_k^l \end{aligned} \tag{5.73}$$

where ϑ_k defines a first-order *decay rate*[1] of species k. If we further assume that the species k could be additionally subjected to other reactions, e.g., $B_m^l + A_k^l \rightarrow \ldots$,

[1] Referring to radioactive decay processes the decay rate ϑ_k is frequently expressed in terms of a reaction *half-life* $t_{1/2k}$ of species k, which is a specific solution of the reaction equation

$$\frac{dC_k}{dt} = -\vartheta_k C_k$$

applied to a simple chemical batch reaction (without diffusion/dispersion and advection). Its analytical solution yields

$$C_k = C_{0k}e^{-\vartheta_k t}$$

The time t for the concentration C_k to decrease from the initial concentration C_{0k} to its half value $\frac{1}{2}C_{0k}$ corresponds to the half-life $t_{1/2}$. From above it results which remain unspecified firstly, we can generalized the reaction rate R_k (5.73) in such a form as

$$\begin{aligned} R_k &= -\varepsilon_l k_A\{A^l_k\} - \varepsilon_{s\text{active}} k_A\{A^s_k\} + \text{`more reaction terms'} \\ &= -\varepsilon s^l \vartheta_k \Re_k C^l_k + \tilde{R}_k \end{aligned} \tag{5.74}$$

where $\tilde{R}_k$ encompasses the additional, yet unspecified reactions of species k. Hence, the mass balance equations (5.68) and (5.69) can be written in the divergence form

$$\frac{\partial}{\partial t}\left(\varepsilon s^l \Re_k C^l_k\right) + \nabla \cdot (\boldsymbol{q}^l C^l_k) + \nabla \cdot \boldsymbol{j}_{lk} + \varepsilon s^l \vartheta_k \Re_k C^l_k = \tilde{R}_k \tag{5.75}$$

and in the convective form

$$\varepsilon s^l \acute{\Re}_k \frac{\partial C^l_k}{\partial t} + \boldsymbol{q}^l \cdot \nabla C^l_k + \nabla \cdot \boldsymbol{j}_{lk} + \varepsilon s^l \vartheta_k \Re_k C^l_k = \tilde{R}_k - C^l_k Q_l \tag{5.76}$$

which obviously become advantageous in their numerical treatment since the linear decay term split off the overall reaction rate appears on the LHS of the balance equations allowing an implicit solution. Thus, formulations in the form of (5.75) and (5.76) are preferably taken into account in the final model equations listed in Tables 3.5, 3.7, and 3.9–3.11 for porous media as well as given by (4.71) and in Table 4.6 for discrete features.

5.4.3 Consecutive Reactions

Considering consecutive reactions (termed also as decay chains or serial reactions, typical in radioactive decay) in the following form

$$A \overset{k_A}{\rightarrow} B \overset{k_B}{\rightarrow} C \overset{k_C}{\rightarrow} D \tag{5.77}$$

the reaction rates for the initial reactant A, the intermediate species B and C as well as the product D (exemplified for the liquid phase l) can be written as

$$t_{1/2k} = \frac{\ln 2}{\vartheta_k}$$

where $\ln 2 = 0.693$ is the natural logarithm of 2. Accordingly, the decay rate ϑ_k can be expressed by

$$\vartheta_k = \frac{\ln 2}{t_{1/2k}}$$

where the half-life $t_{1/2k}$ has to be specified for a given (radioactive) species k.

$$\begin{aligned} R_A &= -\varepsilon_l k_A\{A\} \\ R_B &= \varepsilon_l(k_A\{A\} - k_B\{B\}) \\ R_C &= \varepsilon_l(k_B\{B\} - k_C\{C\}) \\ R_D &= \varepsilon_l k_C\{C\} \end{aligned} \tag{5.78}$$

Additionally, we again assume that such a type of consecutive reaction is subjected to both liquid l and solid s species k for a reversible heterogeneous reaction similar to (5.72), i.e.,

$$\begin{array}{ccccccc} A^l & \overset{k_A}{\rightarrow} & B^l & \overset{k_B}{\rightarrow} & C^l & \overset{k_C}{\rightarrow} & D^l \\ \upharpoonleft\downharpoonright & & \upharpoonleft\downharpoonright & & \upharpoonleft\downharpoonright & & \upharpoonleft\downharpoonright \\ A^s & \overset{k_A}{\rightarrow} & B^s & \overset{k_B}{\rightarrow} & C^s & \overset{k_C}{\rightarrow} & D^s \end{array} \tag{5.79}$$

Using the same procedures as described in Sect. 5.4.2 we find the following set of balance equations for the four species written in the divergence form:

$$\begin{aligned} \tfrac{\partial}{\partial t}\big(\varepsilon s^l \Re_A C_A^l\big) + \nabla\cdot(\boldsymbol{q}^l C_A^l) + \nabla\cdot \boldsymbol{j}_{lA} &= -\varepsilon s^l \vartheta_A \Re_A C_A^l + \tilde{R}_A \\ \tfrac{\partial}{\partial t}\big(\varepsilon s^l \Re_B C_B^l\big) + \nabla\cdot(\boldsymbol{q}^l C_B^l) + \nabla\cdot \boldsymbol{j}_{lB} &= \varepsilon s^l(\vartheta_A \Re_A C_A^l - \vartheta_B \Re_B C_B^l) + \tilde{R}_B \\ \tfrac{\partial}{\partial t}\big(\varepsilon s^l \Re_C C_C^l\big) + \nabla\cdot(\boldsymbol{q}^l C_C^l) + \nabla\cdot \boldsymbol{j}_{lC} &= \varepsilon s^l(\vartheta_B \Re_B C_B^l - \vartheta_C \Re_C C_C^l) + \tilde{R}_C \\ \tfrac{\partial}{\partial t}\big(\varepsilon s^l \Re_D C_D^l\big) + \nabla\cdot(\boldsymbol{q}^l C_D^l) + \nabla\cdot \boldsymbol{j}_{lD} &= \varepsilon s^l \vartheta_C \Re_C C_C^l + \tilde{R}_D \end{aligned} \tag{5.80}$$

In a generalized formulation the equation system (5.80) can concisely be written for species $k = (1 = A), (2 = B), \ldots$, as

$$\frac{\partial}{\partial t}\big(\varepsilon s^l \Re_k C_k^l\big) + \nabla\cdot(\boldsymbol{q}^l C_k^l) + \nabla\cdot \boldsymbol{j}_{lk} = \varepsilon s^l(\vartheta_{k-1}\Re_{k-1}C_{k-1}^l - \vartheta_k \Re_k C_k^l) + \tilde{R}_k \tag{5.81}$$

where it is by definition in (5.81) that $\vartheta_0 = \Re_0 = C_0^l = 0$.

5.4.4 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

$$A + E \underset{k_A^-}{\overset{k_A^+}{\rightleftarrows}} EA \overset{k_B}{\rightarrow} P + E \tag{5.82}$$

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 of *EA* gives as exemplified for the $l-$phase:

$$\begin{aligned} R_{EA} &= \varepsilon_l \big(r_{(k_A^+)} - r_{(k_A^-)} - r_{(k_B)} \big) \\ &= \varepsilon_l \big(k_A^+ \{A\}\{E\} - k_A^- \{EA\} + k_B \{EA\} \big) = 0 \end{aligned} \tag{5.83}$$

It follows

$$\{EA\} = \frac{\{A\}\{E\}}{k_m}, \qquad k_m = \frac{k_A^- + k_B}{k_A^+} \tag{5.84}$$

Introducing the total concentration of enzyme as

$$\{E_T\} = \{E\} + \{EA\} \tag{5.85}$$

and assuming only a little enzyme is added so that $\{A\}$ differs only slightly from its total concentration, then

$$\{EA\} = \frac{\{A\}\big(\{E_T\} - \{EA\}\big)}{k_m} \tag{5.86}$$

which rearranges to

$$\{EA\} = \frac{\{E_T\}\{A\}}{k_m + \{A\}} \tag{5.87}$$

On the other hand, the reaction rate for species A in the $l-$phase is

$$R_A = -\varepsilon_l \big(k_A^+ \{A\}\{E\} - k_A^- \{EA\} \big) \tag{5.88}$$

which can be simplified by applying the pre-equilibrium condition (5.83) as

$$R_A = -\varepsilon_l k_B \{EA\} \tag{5.89}$$

Inserting (5.87) into (5.89) it yields finally

$$R_A = -\varepsilon_l \frac{k_B \{E_T\}\{A\}}{k_m + \{A\}} = -\varepsilon_l \frac{v_m C_A^l}{K_m + C_A^l} \tag{5.90}$$

where

$$\begin{aligned} v_m &= k_B \{E_T\} && \text{maximum velocity of enzymolysis} \\ K_m &= \frac{m_A k_m}{\gamma_A} && \text{Michaelis (Monod) constant} \end{aligned} \tag{5.91}$$

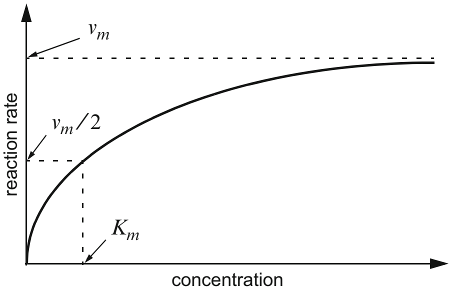


Fig. 5.1 Saturation curve of the Michaelis-Menten reaction rate

The intrinsic reaction rate $\hat{R}_A = |R_A|/\varepsilon_l = v_m C_A^l/(K_m + C_A^l)$ of (5.90) results in a saturation curve[2] as shown in Fig. 5.1, where v_m appears as the *maximum growth rate* and K_m as the *half-saturation constant*.

Let us generalize the reaction (5.82) if we assume that the Michaelis-Menten kinetics is subjected to both the liquid and solid species k of a reversible reaction similar to (5.72). In doing so, we are interested in a reaction of the type:

$$\begin{aligned} |\nu_1| A_k^l + |\nu_2| A_m^s &\rightleftharpoons |\nu_1| A_k^s + |\nu_2| A_m^l \\ A_k^l + E^l &\underset{k_A^-}{\overset{k_A^+}{\rightleftarrows}} EA^l \overset{k_B}{\rightarrow} P^l + E^l \\ A_k^s + E^s &\underset{k_A^-}{\overset{k_A^+}{\rightleftarrows}} EA^s \overset{k_B}{\rightarrow} P^s + E^s \end{aligned} \tag{5.92}$$

which leads to the following reaction rate of species k

[2] Considering the Michaelis-Menten reaction rate in the form $\hat{R}_A = v_m C_A^l/(K_m + C_A^l)$:

(i) If C_A^l is large compared to K_m then $C_A^l/(K_m + C_A^l) \approx 1$ and the reaction rate becomes

$$\hat{R}_A \approx v_m$$

(ii) If $C_A^l = K_m$ then $C_A^l/(K_m + C_A^l) = \frac{1}{2}$ and the reaction rate gives

$$\hat{R}_A = \tfrac{1}{2} v_m$$

(iii) If If C_A^l is small compared to K_m then $C_A^l/(K_m + C_A^l) \approx C_k^l/K_m$ and it is

$$\hat{R}_A = \frac{v_m}{K_m} C_k^l$$

$$
\begin{aligned}
R_k &= -\varepsilon_l \frac{v_m C_k^l}{K_m + C_k^l} - \varepsilon_{s^{\text{active}}} \frac{v_m C_k^s}{K_m + C_k^s} \\
&= -s^l \left[\varepsilon \frac{v_m C_k^l}{K_m + C_k^l} + (1-\varepsilon) \frac{v_m \varphi_k C_k^l}{K_m + \varphi_k C_k^l} \right] \\
&= -\varepsilon s^l \underbrace{\left(\frac{v_m}{K_m + \varphi_k C_k^l} \right)}_{\vartheta_k^m} \left[1 + \frac{(1-\varepsilon)}{\varepsilon} \varphi_k + \underbrace{\left(\frac{\varphi_k - 1}{1 + \frac{K_m}{C_k^l}} \right)}_{\zeta_k^m} \right] C_k^l \\
&= -\varepsilon s^l \vartheta_k^m (\Re_k + \zeta_k^m) C_k^l
\end{aligned}
\tag{5.93}
$$

where $\vartheta_k^m = v_m/(K_m + \varphi_k C_k^l)$ defines a specific Michaelis-Menten 'decay rate' and $\zeta_k^m = (\varphi_k - 1)/(1 + \frac{K_m}{C_k^l})$ is a modifying function. Inserting the reaction rate (5.93) into (5.68) the following mass transport equation results

$$
\frac{\partial}{\partial t}\left(\varepsilon s^l \Re_k C_k^l\right) + \nabla \cdot (\boldsymbol{q}^l C_k^l) + \nabla \cdot \boldsymbol{j}_{lk} = -\varepsilon s^l \vartheta_k^m (\Re_k + \zeta_k^m) C_k^l \tag{5.94}
$$

where the retardation factor $\Re_k$ is defined by (5.70). We note that ϑ_k^m and ζ_k^m are functions of C_k^l.

5.5 Generalized Kinetic Formulations

In the preceding Sects. 5.3 and 5.4 the bulk reaction rate R_k of species k

$$
R_k = \sum_\alpha \varepsilon_\alpha (r_k^\alpha + R_k^\alpha) \qquad (\alpha = l, s), (k = 1, \ldots, N) \tag{5.95}
$$

consisting of the parts of homogeneous and heterogeneous reactions r_k^α and R_k^α, respectively, and the deduced bulk reaction rate $\tilde{R}_k$

$$
\tilde{R}_k = R_k + \sum_\alpha \varepsilon_\alpha \vartheta_k C_k^\alpha \qquad (\alpha = l, s), (k = 1, \ldots, N) \tag{5.96}
$$

separated by a linear decay reaction controlled via the decay constant ϑ_k (we note that $\tilde{R}_k = R_k$ if we use $\vartheta_k = 0$), could be developed by kinetic formulations in dependence on the reaction type and stoichiometry. Reactive systems can be broadly classified into simple and complex kinetic systems. The former consists of elementary unimolecular and bimolecular reactions while the latter encompasses opposing, concurrent and consecutive reactions. The progress of a reaction can be limited by the availability of reactants or intermediate species, and may be slowed not only by the presence of reaction products, but also by other inhibiting species. In addition, the reaction progress may be catalyzed by species, which are not directly

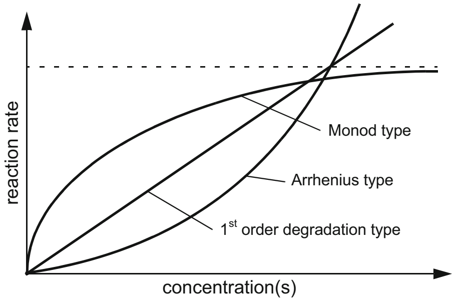


Fig. 5.2 Variation of reaction rate with concentration

involved in the reaction. According to the mechanism of a given reaction, the functional form of R_k (or $\tilde{R}_k$) can be very complicated and requires a more general approach to make the model applicable to a wide range of problems subjected to kinematically controlled reactions. Laboratory experiments have to be conducted to determine which species control the reaction and what order the reaction has with respect to each of these species.

A typical *constitutive representation* of R_k (or $\tilde{R}_k$) has a functional

$$R_k = R_k(C_1^\alpha, C_2^\alpha, \dots, C_N^\alpha, \varepsilon_\alpha, T) \qquad (\alpha = l, s), (k = 1, \dots, N) \tag{5.97}$$

which can be developed by a polynomial expression of low order in terms of concentrations C_k^α for simple kinetic systems or more complex rate expressions of higher order, cf. Fig. 5.2. We can distinguish the following classes of rate expressions.

5.5.1 Degradation Type Kinetics

The term 'degradation' is loosely used in the literature and refers to some measure of mass loss or change in species concentration over time. For a degradation type kinetics the reaction rate R_k of species k can be developed in a polynomial representation of low order, viz.,

$$\begin{aligned} R_k &= \varepsilon_{\alpha_1} \nu_1 k_1 \left(C_1^\alpha\right)^{n_1} + \varepsilon_{\alpha_2} \nu_2 k_2 \left(C_2^\alpha\right)^{n_2} + \dots + \varepsilon_{\alpha_N} \nu_N k_N \left(C_N^\alpha\right)^{n_N} \\ &= \sum_{m=1}^{N} \left[\varepsilon_{\alpha_m} \nu_m k_m \left(C_m^\alpha\right)^{n_m} \right] \qquad (k = 1, \dots, N) \end{aligned} \tag{5.98}$$

where ε_{α_m} indicates the volume fraction of the phase α containing the species m, ν_m $(m = 1, \dots, N)$ are stoichiometric coefficients controlling the signs of the reaction terms ($\nu_m < 0$ for reactants and $\nu_m > 0$ for products) and can additionally weight the rate constants k_m $(m = 1, \dots, N)$. The rate constants k_m can be dependent on the temperature T. In (5.98) $n_m \geq 0$ represents the exponent of species m. In the case of $n_m = 1, \ \forall m$ a 1st-order degradation type results. We note that the decay reactions as described in Sects. 5.4.2 and 5.4.3 belong to this type of degradation kinetics, which can be written as

$$R_k = \sum_{m=1}^{N} \varepsilon_{\alpha_m} \nu_m k_m C_m^\alpha \qquad (k = 1, \dots, N) \tag{5.99}$$

or

$$\tilde{R}_k = \sum_{\substack{m=1 \\ m \neq k}}^{N} \varepsilon_{\alpha_m} \nu_m k_m C_m^\alpha \qquad (k = 1, \dots, N) \tag{5.100}$$

A typical example is the radionuclide decay chain of uranium (5.32), where the reaction rate for the radium species ($k = \text{Ra}$) of the liquid phase l becomes the form

$$R_{\text{Ra}} = \varepsilon_l (k_{\text{Th}} C_{\text{Th}}^l - k_{\text{Ra}} C_{\text{Ra}}^l) \tag{5.101}$$

where k_{Th} and k_{Ra} are known rate constants (decay rates) of thorium and radium, respectively.

5.5.2 Arrhenius Type Kinetics

The Arrhenius type kinetics is expressed by a polynomial representation of higher order for the reaction rate R_k written in the form

$$\begin{aligned} R_k &= \varepsilon_{\alpha_1} \nu_1 k_1 \prod_{n=1}^{N^o} (C_n^\alpha)^{n_n} + \dots + \varepsilon_{\alpha_N} \nu_N k_N \prod_{n=1}^{N^o} (C_n^\alpha)^{n_n} \\ &= \sum_{m=1}^{N} \left[\varepsilon_{\alpha_m} \nu_m k_m \prod_{n=1}^{N^o} (C_n^\alpha)^{n_n} \right] \qquad (k = 1, \dots, N) \end{aligned} \tag{5.102}$$

where the reaction constants $k_m \ \ (m = 1, \dots, N)$ are given by the *Arrhenius equation* as

$$k_m = A_m \exp\Big(-\frac{E^\#}{RT}\Big) \qquad (m = 1, \dots, N) \tag{5.103}$$

in which A_m is the pre-exponential factor, R is the molar gas constant (or universal thermodynamic constant (∼8.314 J/°K mole)), T is the absolute temperature and $E^\#$ is the activation energy to be known [13]. The two quantities A_m and $E^\#$ are commonly termed as *Arrhenius parameters.*

A typical example of an Arrhenius reaction is the uraninite leaching (5.31), where we can find the following ration rates of the six species [$k = 1, \ldots, 6: 1 = \mathrm{UO_2}(s), 2 = \mathrm{O_2}, 3 = \mathrm{CO_3^{2-}}, 4 = \mathrm{HCO_3^-}, 5 = \mathrm{UO_2(CO_3)_3^{4-}}$ and $6 = \mathrm{H_2O}$ ($N = 6$)] subjected to four reactants [$\mathrm{UO_2}(s), \mathrm{O_2}, \mathrm{CO_3^{2-}}$ and $\mathrm{HCO_3^-}$ ($N^o = 4$)] involved in the process, e.g., [386]:

$$\begin{aligned} R_1 &= -\varepsilon_s A_1 \exp\bigl(-\tfrac{E^\#}{RT}\bigr) C_1^s C_2^l C_3^l C_4^l \\ R_2 &= -\tfrac{1}{2}\, \varepsilon_l A_2 \exp\bigl(-\tfrac{E^\#}{RT}\bigr) C_1^s C_2^l C_3^l C_4^l \\ R_3 &= -\varepsilon_l A_3 \exp\bigl(-\tfrac{E^\#}{RT}\bigr) C_1^s C_2^l C_3^l C_4^l \\ R_4 &= -2\, \varepsilon_l A_4 \exp\bigl(-\tfrac{E^\#}{RT}\bigr) C_1^s C_2^l C_3^l C_4^l \\ R_5 &= \varepsilon_l A_5 \exp\bigl(-\tfrac{E^\#}{RT}\bigr) C_1^s C_2^l C_3^l C_4^l \\ R_6 &= \varepsilon_l A_6 \exp\bigl(-\tfrac{E^\#}{RT}\bigr) C_1^s C_2^l C_3^l C_4^l \end{aligned} \tag{5.104}$$

where A_k ($k = 1, \ldots, 6$), $E^\#$ and R have to be known parameters. Note that the stoichiometric coefficients for the uraninite leaching reaction (5.31) are: $\nu_1 = -1, \nu_2 = -1/2, \nu_3 = -1, \nu_4 = -2, \nu_5 = 1, \nu_6 = 1$. Furthermore it is $\varepsilon_{\alpha_1} = \varepsilon_s$ and $\varepsilon_{\alpha_2} = \varepsilon_{\alpha_3} = \varepsilon_{\alpha_4} = \varepsilon_{\alpha_5} = \varepsilon_{\alpha_6} = \varepsilon_l$.

5.5.3 Monod Type Kinetics

Monod type kinetics describe more complex biochemical reaction systems. Monod was the first to recognize that the growth rate of a microbial population is restricted by the concentration of the growth-limiting substrate. Monod established that the form of the relationship was analogous to the Michaelis-Menten enzyme kinetics equation, see Sect. 5.4.4. The Monod kinetics can be extended by inhibition parameters. Its mathematical representation can be written in a generalized form as

$$R_k = \varepsilon_{\alpha_k} \nu_k \left(\frac{v_k C_k^\alpha}{K_k + C_k^\alpha} \right) \left[\left(C_{m \neq k}^\alpha \right)^{n_m} \prod_{\substack{n \neq k \\ n \neq m}}^{N^o} \left(\frac{k_n^* \left(C_n^\alpha \right)^{p_n}}{K_n^* + \left(C_n^\alpha \right)^{n_n}} \right) \right] \tag{5.105}$$

where v_k is the maximum growth rate, K_k is the half-saturation constant of species k, k_n^* and K_n^* are inhibition coefficients related to species n. The exponents p_n and n_n can be independently chosen so that the concentration dependency can be reduced in case of need if setting for instance $p_n = 0$ or $n_n = 0$ and so forth.

In (5.105) C_k^α represents the concentration of the growth-limiting substrate. The half-saturation constant K_k in a biochemical context may be viewed as a measure of the affinity the microorganisms have for the growth-limiting substrate: (1) the lower the value of K_k, the greater the capacity to grow rapidly in an environment with low concentrations of growth-limiting substrate, and (2) the lower the value of K_k, the lower the growth-limiting substrate concentration at which the maximum specific growth rate v_k is attained (see also footnote of Sect. 5.4.4).

A typical example of a Monod kinetics is the biodegradation of BTEX (5.30), where six species ($k = 1, \ldots, 6$) are involved: (oxygen – 1 = O_2, nitrate – 2 = NO_3, iron – 3 = Fe^{2+}, sulfate – 4 = SO_4, methane – 5 = CH_4 and hydrocarbon – 6 = HC). The following reaction rates can be given [86]:

$$\begin{aligned} R_1 &= \varepsilon_l \nu_1\, r_1 \\ R_2 &= \varepsilon_l \nu_2\, r_2 \\ R_3 &= \varepsilon_l \nu_3\, r_3 \\ R_4 &= \varepsilon_l \nu_4\, r_4 \\ R_5 &= \varepsilon_l \nu_5\, r_5 \\ R_6 &= -\varepsilon_l \textstyle\sum_{m=1}^{5} r_m \end{aligned} \tag{5.106}$$

with

$$\nu_1 = -3.14, \quad \nu_2 = -4.9, \quad \nu_3 = 21.8, \quad \nu_4 = -4.7, \quad \nu_5 = 0.78 \tag{5.107}$$

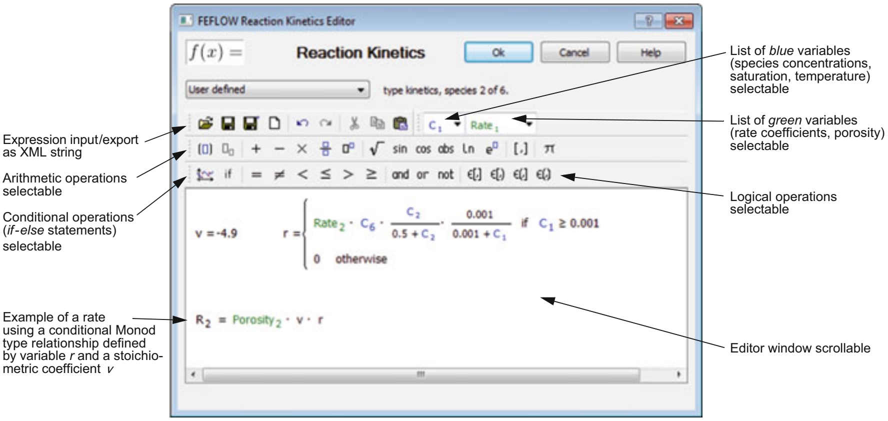
and the specific rates

$$\begin{aligned} r_1 &= C_6^l \Big(\frac{v_1 C_1^l}{K_1 + C_1^l}\Big) \\ r_2 &= C_6^l \Big(\frac{v_2 C_2^l}{K_2 + C_2^l}\Big)\Big(\frac{K_1^*}{K_1^* + C_1^l}\Big) \\ r_3 &= C_6^l \Big(\frac{v_3 \check{C}_3^l}{K_3 + \check{C}_3^l}\Big)\Big(\frac{K_1^*}{K_1^* + C_1^l}\Big)\Big(\frac{K_2^*}{K_2^* + C_2^l}\Big) \\ r_4 &= C_6^l \Big(\frac{v_4 C_4^l}{K_4 + C_4^l}\Big)\Big(\frac{K_1^*}{K_1^* + C_1^l}\Big)\Big(\frac{K_2^*}{K_2^* + C_2^l}\Big)\Big(\frac{K_3^*}{K_3^* + \check{C}_3^l}\Big) \\ r_5 &= C_6^l \Big(\frac{v_5 \check{C}_5^l}{K_5 + \check{C}_5^l}\Big)\Big(\frac{K_1^*}{K_1^* + C_1^l}\Big)\Big(\frac{K_2^*}{K_2^* + C_2^l}\Big)\Big(\frac{K_3^*}{K_3^* + C_3^l}\Big)\Big(\frac{K_4^*}{K_4^* + C_4^l}\Big) \end{aligned} \tag{5.108}$$

in which $\check{C}_m^l = C_m^{l\max} - C_m^l$, where $C_m^{l\max}$ are known (measurable) maximum concentrations of a species m.

5.5.4 *Freely Editable Kinetic Expressions of R_k*

While the reaction mechanisms categorized above as reactions of degradation (5.98), Arrhenius (5.102) and Monod (5.105) cover a wide spectrum for practical applications, a number of chemical reactions, however, need more degrees of freedom to specify their kinetic relationships in a higher complexity and structure. For these purposes FEFLOW provides a *reaction kinetics editor*, where the reaction



User defined

type kinetics, species 2 of 6.

$v = -4.9$

$r = \begin{cases} Rate_2 \cdot C_6 \cdot \frac{C_2}{0.5 + C_2} \cdot \frac{0.001}{0.001 + C_1} & \text{if } C_1 \geq 0.001 \\ 0 & \text{otherwise} \end{cases}$

$R_2 = Porosity_2 \cdot v \cdot r$

Expression input/export as XML string

Arithmetic operations selectable

Conditional operations (*if-else* statements) selectable

Example of a rate using a conditional Monod type relationship defined by variable *r* and a stoichiometric coefficient *v*

List of *blue* variables (species concentrations, saturation, temperature) selectable

List of *green* variables (rate coefficients, porosity) selectable

Logical operations selectable

Editor window scrollable

Fig. 5.3 FEFLOW's reaction kinetics editor

rates R_k can be freely edited in a graphical and interactive manner without any limitations on the algebraic structure of the rate expressions. The rate expressions of FEFLOW's reaction kinetics editor do not require any programming and code compiling. A fast code interpreter provides efficient computations of reaction expressions during the numerical simulations. The advantages of this reaction kinetics editor are in particular:

- Freely editable algebraic expressions of R_k in an arbitrary structure and complexity,
- Combination with conditional *if-else* statements, and
- Including of depending variables (i.e., access to all concentrations C_k^α $(k = 1, \ldots, N)$, $(\alpha = l, s)$ participating in the reaction system, moreover, to saturation s^l and temperature T in case of need) and spatially and/or temporarily variable material parameters (e.g., porosity ε and solid fraction ε_s) in the rate expressions of R_k.

The dialog of FEFLOW's reaction kinetics editor is shown in Fig. 5.3, where solution variables (e.g., all concentrations C_k^α $(k = 1, \ldots, N)$, $(\alpha = l, s)$, liquid saturation s^l or temperature T) appear as *blue* entities and variable parameters (e.g., reaction rates $k_k, \ldots$, porosity ε or solid fraction ε_s) appear as *green* entities to compose the reaction rate expressions. Arithmetic and logical operations can be included in the formulations.

Chapter 6
Initial, Boundary and Constraint Conditions

6.1 Introduction

The governing model equations derived in Chaps. 3 and 4, which are summarized in Table 3.7 for general variably saturated porous media, in Table 3.9 for fully saturated porous media (groundwater), in Table 3.10 for 2D unconfined aquifers and in Table 3.11 for 2D confined aquifers as well as in Tables 4.5–4.7 for variable-density flow, mass and heat transport of discrete features, have to be supplemented by initial, boundary and constraint conditions. The solutions for the flow, mass and heat transport equations are generally sought within a domain $\Omega \subset \Re^D$ closed by its boundary $\Gamma \subset \Re^D$ $(D = 1, 2, 3)$ in the $D-$dimensional Euclidean space (cf. Sect. 2.2.2). By definition, the boundary Γ is separated from the domain Ω. On the other hand, by $\bar{\Omega}$ we denote the (closure) domain, which completely joins the boundary

$$\bar{\Omega} = \Omega \cup \Gamma \tag{6.1}$$

On $\bar{\Omega}$ and Γ initial conditions (IC's) and boundary conditions (BC's) have to be specified, respectively. The boundary Γ consists of disjoint nonoverlapping portions Γ_i $(i = 1, 2, \ldots)$ bounding the domain Ω both outside and inside, which can be suitably subdivided according to the types of BC's. BC's are always required for both transient and steady-state problems, while IC's are always needed for transient problems. An exception possesses nonlinear steady-state problems, where an IC of the solution initializes an iterative procedure.

In addition, singular point conditions (SPC's) are of interest for specifying pumping (discharging) or injection (recharging) wells, which are assigned to separate points of the domain Ω. Due to the nature of singularities well-type SPC's must be treated in a singular (discrete) manner which is different to the treatment of BC's, where fluxes are continuous and integrable over a boundary section. It is interesting to note that the effect by a flux-type BC can be similar or even identical

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_6,

to a SPC specification applied to a numerical model for cases, where all connected points forming the discretized boundary are imposed by a respective SPC.

6.2 Initial Conditions (IC's)

In the domain $\bar{\varOmega}$ the following IC's are valid for the flow, species mass and heat transport processes, respectively:
Flow

$$h(\boldsymbol{x}, t_0) = h_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{6.2}$$

Mass transport of species k

$$C_k(\boldsymbol{x}, t_0) = C_{k0}(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{6.3}$$

Heat transport

$$T(\boldsymbol{x}, t_0) = T_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{6.4}$$

where h_0, C_{k0} and T_0 are known spatially varying functions of initial distribution at initial time t_0.

6.3 Standard Boundary Conditions (BC's) and Well-Type Singular Point Conditions (SPC's)

On the boundary $\varGamma$ closing the domain $\varOmega$ disjoint portions are appropriately defined as $\varGamma_i$ $(i = 1, 2, \ldots)$ for which different types of BC can be separately specified. Dirichlet-type (1st kind or *essential*) BC's on $\varGamma_1$, $\varGamma_4$ and $\varGamma_7$, Neumann-type (2nd kind) BC's on $\varGamma_2$, $\varGamma_5$ and $\varGamma_8$ as well as Cauchy (Robin)-type (3rd kind) BC's on $\varGamma_3$, $\varGamma_6$ and $\varGamma_9$ will represent standard formulations (cf. Sect. 2.2.2) for flow, mass and heat, respectively, so that for standard BC's: $\varGamma = \varGamma_1 \cup \varGamma_2 \cup \varGamma_3 = \varGamma_4 \cup \varGamma_5 \cup \varGamma_6 = \varGamma_7 \cup \varGamma_8 \cup \varGamma_9$. Additionally, well-type SPC's are included providing specific sink/source conditions which have to be assigned to separate points of the domain $\varOmega$ idealized as wells. Furthermore, integrated formulations of Neumann-type and Cauchy-type BC's are desired. BC's of 1st, 2nd and 3rd will be symbolized by ○, × and ⊗, respectively. A well-type SPC will be symbolized by ↱. Special formulations of BC's are necessary in various applications which will be introduced in Sect. 6.5 further below.

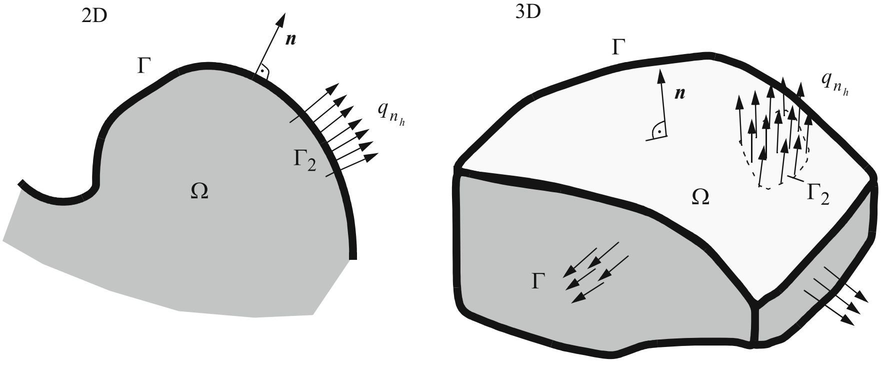


Fig. 6.1 Normal Neumann-type fluxes for 2D and 3D boundary geometries

6.3.1 Flow BC

6.3.1.1 ○ Dirichlet-Type (1st Kind) BC

$$h(\boldsymbol{x},t) = h_D(t) \quad \text{on} \quad \varGamma_1 \times t[t_0,\infty) \tag{6.5}$$

where h_D are prescribed values of hydraulic head on $\varGamma_1 \subset \varGamma$. Note that for steady-state flow problems Dirichlet-type BC's (6.5) are usually required, i.e., $\varGamma_1 \neq \emptyset$, unless Cauchy-type BC's occur.

6.3.1.2 × Neumann-Type (2nd Kind) BC

$$\left.\begin{array}{ll} q_{n_h}(\boldsymbol{x},t) = -\big[k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})\big] \cdot \boldsymbol{n} & = q_h(t) \\ \qquad\qquad \text{for 3D and 2D vertical \& unconfined} & \\ \bar{q}_{n_h}(\boldsymbol{x},t) = -(\boldsymbol{T} f_\mu \cdot \nabla h) \cdot \boldsymbol{n} & = \bar{q}_h(t) \\ \qquad\qquad \text{for 2D horizontal, confined} & \end{array}\right\} \quad \text{on} \quad \varGamma_2 \times t[t_0,\infty) \tag{6.6}$$

where $\boldsymbol{n}$ is the positive outward-directed unit normal to $\varGamma_2$, $q_{n_h} = \boldsymbol{q} \cdot \boldsymbol{n}$ and $\bar{q}_{n_h} = \bar{\boldsymbol{q}} \cdot \boldsymbol{n}$ represent normal fluxes (positive outward-directed) across the boundary $\varGamma_2$ and q_h and $\bar{q}_h$ are the prescribed Neumann fluxes on $\varGamma_2 \subset \varGamma$ as illustrated in Fig. 6.1. If $q_h = 0$ and $\bar{q}_h = 0$ the Neumann-type BC reduces to a *natural* (no-flux) BC associated with $\nabla h + \chi \boldsymbol{e} = \boldsymbol{0}$ and $\nabla h = \boldsymbol{0}$, respectively. Note that for saturated porous media $k_r = 1$, for density-uncoupled problems $\chi = 0$ and for constant liquid viscosity, equal to the reference viscosity, $f_\mu = 1$. For 2D horizontal unconfined aquifer problems with $k_r = 1$ and $\chi = 0$, the prescribed Neumann flux q_h has to be vertically integrated in accordance with the unknown water table h.

6.3.1.3 ⊗ Cauchy-Type (3rd Kind) BC

$$\left.\begin{aligned} q_{n_h}(\boldsymbol{x},t) &= -\big[k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})\big]\cdot \boldsymbol{n} &&= -\varPhi_h(h_C - h) \\ &\quad \text{for 3D and 2D vertical \& unconfined} && \\ \bar{q}_{n_h}(\boldsymbol{x},t) &= -(\boldsymbol{T} f_\mu \cdot \nabla h)\cdot\boldsymbol{n} &&= -\bar{\varPhi}_h(h_C-h) \\ &\quad \text{for 2D horizontal, confined} && \end{aligned}\right\} \text{ on } \quad \varGamma_3 \times t[t_0,\infty) \tag{6.7}$$

where h_C are prescribed values of hydraulic head on $\varGamma_3 \subset \varGamma$. The signs of $q_{n_h} = \boldsymbol{q}\cdot\boldsymbol{n}$ and $\bar{q}_{n_h} = \bar{\boldsymbol{q}} \cdot \boldsymbol{n}$ are chosen that the boundary fluxes are positive outward-directed if $h > h_C$. In (6.7) the *transfer coefficients* $\varPhi_h$ and $\bar{\varPhi}_h$ represent *dual* directional functions in form of:

$$\varPhi_h = \begin{cases} \varPhi_h^{\text{in}}(\boldsymbol{x},t) & \text{for} \quad h_C > h \\ \varPhi_h^{\text{out}}(\boldsymbol{x},t) & \text{for} \quad h_C \le h \end{cases} \tag{6.8}$$

$$\bar{\varPhi}_h = \begin{cases} \bar{\varPhi}_h^{\text{in}}(\boldsymbol{x},t) & \text{for} \quad h_C > h \\ \bar{\varPhi}_h^{\text{out}}(\boldsymbol{x},t) & \text{for} \quad h_C \le h \end{cases} \tag{6.9}$$

which are in general functions of space $\boldsymbol{x}$ and time t. Accordingly, in specifying two alternate (if necessary temporal) transfer coefficients different transfer conditions can be input to distinguish between inflow conditions ($q_{n_h} < 0$, e.g., infiltration from a surface water into the aquifer) and outflow conditions ($q_{n_h} > 0$, e.g., exfiltrating the aquifer into the surface water). Their usefulness for river-aquifer interactions is discussed further below. The special case $\varPhi_h = \varPhi_h^{\text{in}} = \varPhi_h^{\text{out}}$ or $\bar{\varPhi}_h = \bar{\varPhi}_h^{\text{in}} = \bar{\varPhi}_h^{\text{out}}$ does not differ between inward and outward boundary flux, so it becomes directionally independent.

The formulation of 3rd kind BC's is based on a general transfer relation between the reference value h_C on the boundary portion $\varGamma_3$ and the hydraulic head h to be computed at the same place. The reference hydraulic head h_C can also be time-dependent $h_C = h_C(t)$. The dual transfer coefficient $\varPhi_h$ possesses the property of a resistance coefficient which constrains the discharge through the boundary and, additionally, differs between inflow and outflow conditions by means of $\varPhi_h^{\text{in}}$ and $\varPhi_h^{\text{out}}$, respectively, according to (6.8) and (6.9). If $\varPhi_h \equiv 0$ the boundary becomes impervious. On the other hand, using a very large value $\varPhi_h \to \infty$ the BC of 3rd kind is reduced to a Dirichlet-type (1st kind) BC approaching to $h = h_C$ on $\varGamma_3$.

For flow problems the transfer coefficient $\varPhi_h$ can be identified as a specific *colmation* (or *leakage*) coefficient as outlined in Fig. 6.2 for inflow (infiltration) conditions ($\varPhi_h \to \varPhi_h^{\text{in}}$ $(h_C > h)$). An adjacent river bed is clogged ('colmated') by a layer of thickness d and a hydraulic conductivity of K_o^{in}. Commonly, the layer conductivity K_o^{in} is much smaller than the conductivity K_1 of the aquifer to be modeled. Thereby the model boundary $\varGamma$ represents the inner boundary of the 'colmation' layer $\varGamma_3$, where the model domain $\varOmega$ ends.

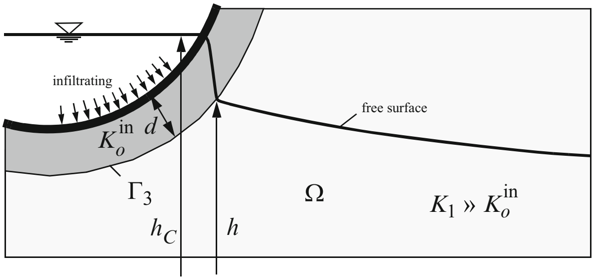


Fig. 6.2 Transfer coefficient $\Phi_h (= \Phi_h^{\text{in}})$ as 'colmation' parameter of a clogged river bed

The flux through such a 'colmation' layer can be estimated from the Darcy equation (see Fig. 6.2), viz.,

$$q_{n_h} \approx -K_o^{\text{in}} \frac{\Delta h}{\Delta s} = -K_o^{\text{in}} \frac{h_C - h}{d} \tag{6.10}$$

where s and Δs identify the arc length and line distance in direction of flow, respectively. Setting (6.7) equal to (6.10) a simple relationship results for the transfer coefficient Φ_h^{in} in 3D and 2D (vertical, horizontal unconfined) cases:

$$\Phi_h^{\text{in}} = \frac{K_o^{\text{in}}}{d} \tag{6.11}$$

For horizontal confined flow problems an inherent vertical averaging becomes necessary (in the aquifer all fluxes are integrated over the depth) resulting in a depth-integrated transfer coefficient $\bar{\Phi}_h^{\text{in}}$ as:

$$\bar{\Phi}_h^{\text{in}} = B\Phi_h^{\text{in}} = B\frac{K_o^{\text{in}}}{d} \tag{6.12}$$

For outward directed (exfiltrating) boundary fluxes according to Fig. 6.3 the following relationships for Φ_h^{out} and $\bar{\Phi}_h^{\text{out}}$ can be derived, analogously to the above, viz.,

$$\Phi_h^{\text{out}} = \frac{K_o^{\text{out}}}{d} \tag{6.13}$$

$$\bar{\Phi}_h^{\text{out}} = B\Phi_h^{\text{out}} = B\frac{K_o^{\text{out}}}{d} \tag{6.14}$$

The coefficients Φ_h^{in} and Φ_h^{out} (also $\bar{\Phi}_h^{\text{in}}$ and $\bar{\Phi}_h^{\text{out}}$) differ if in case of infiltration the conductivities of the 'colmation' layer become depart from that of the exfiltration $K_o^{\text{in}} \neq K_o^{\text{out}}$.

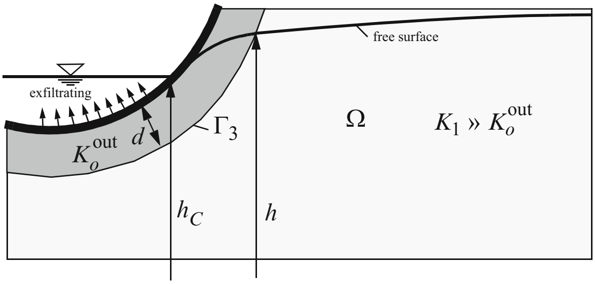


Fig. 6.3 Transfer coefficient $\Phi_h (= \Phi_h^{\text{out}})$ as 'colmation' parameter of a clogged river bed

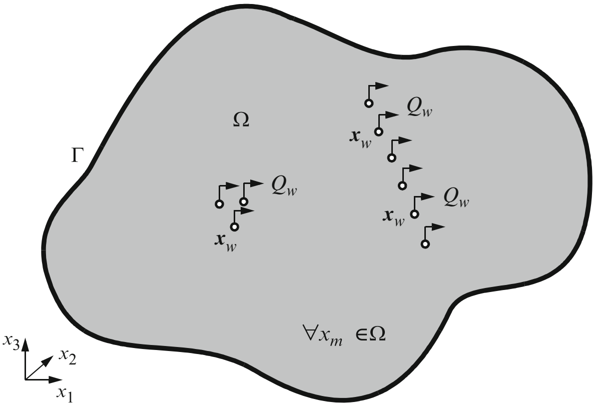


Fig. 6.4 Number of well-type SPC's on singular points $\forall \boldsymbol{x}_m \in \Omega$

6.3.1.4 ↱ Well-Type SPC

A number of pumping (or injecting) wells are idealized as singular point sinks (or sources) at locations $\boldsymbol{x}_w \in \Omega$ (Fig. 6.4):

$$Q_{hw}(\boldsymbol{x}, t) = -\sum_{w=1}^{N_\text{W}} Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) \quad \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty) \tag{6.15}$$

where Q_{hw} is the specific sink/source function of wells, N_W is the number of wells, $Q_w(t)$ is the prescribed volume per unit time discharge (pumping rate) of single well w at location $\boldsymbol{x}_w$ and $\delta(\boldsymbol{x} - \boldsymbol{x}_w) = \prod_{i=1}^{D} \delta(x_i - x_{iw})$ is the Dirac delta function associated with location $\boldsymbol{x}_w$. The Dirac delta $\delta(\boldsymbol{x} - \boldsymbol{x}_w)$ is zero at all points except $\boldsymbol{x} = \boldsymbol{x}_w$ and satisfies

$$\int_\Omega \delta(\boldsymbol{x} - \boldsymbol{x}_w) d\Omega = 1 \tag{6.16}$$

and accordingly

$$\int_\Omega \sum_{w=1}^{N_\text{W}} Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) d\Omega = \sum_{w=1}^{N_\text{W}} Q_w(t) \tag{6.17}$$

6.3.2 Mass Transport BC

6.3.2.1 ○ Dirichlet-Type (1st Kind) BC

$$C_k(\boldsymbol{x},t) = C_{kD}(t) \quad \text{on} \quad \mathit{\Gamma}_4 \times t[t_0,\infty) \tag{6.18}$$

where C_{kD} are prescribed values of concentration of species k on $\mathit{\Gamma}_4 \subset \mathit{\Gamma}$. Note that for steady-state mass transport problems Dirichlet-type BC's (6.18) are usually required, i.e., $\mathit{\Gamma}_4 \neq \emptyset$, unless Cauchy-type BC's occur.

6.3.2.2 × Neumann-Type (2nd Kind) BC

For 3D and 2D (vertical and axisymmetric):

$$\left.\begin{aligned} &\text{convective form}\\ &q_{n_{kC}}(\boldsymbol{x},t) = \underbrace{-(\boldsymbol{D}_k \cdot \nabla C_k)\cdot \boldsymbol{n}}_{\text{dispersive flux}} \qquad\qquad = q_{kC}(t)\\ &\text{divergence form}\\ &q_{n_{kC}}(\boldsymbol{x},t) = \underbrace{C_k\, q_{n_h} - (\boldsymbol{D}_k \cdot \nabla C_k)\cdot \boldsymbol{n}}_{\text{total flux}} = q^{\dagger}_{kC}(t) \end{aligned}\right\} \quad \text{on} \quad \mathit{\Gamma}_5 \times t[t_0,\infty) \tag{6.19}$$

and for 2D horizontal (confined and unconfined):

$$\left.\begin{aligned} &\text{convective form}\\ &\bar{q}_{n_{kC}}(\boldsymbol{x},t) = \underbrace{-(\bar{\boldsymbol{D}}_k \cdot \nabla C_k)\cdot \boldsymbol{n}}_{\text{dispersive flux}} \qquad\qquad = \bar{q}_{kC}(t)\\ &\text{divergence form}\\ &\bar{q}_{n_{kC}}(\boldsymbol{x},t) = \underbrace{C_k\, \bar{q}_{n_h} - (\bar{\boldsymbol{D}}_k \cdot \nabla C_k)\cdot \boldsymbol{n}}_{\text{total flux}} = \bar{q}^{\dagger}_{kC}(t) \end{aligned}\right\} \quad \text{on} \quad \mathit{\Gamma}_5 \times t[t_0,\infty) \tag{6.20}$$

where $q_{n_{kC}}$ and $\bar{q}_{n_{kC}}$ represent normal mass fluxes of species k (positive outward-directed) across the boundary $\mathit{\Gamma}_5$ and q_{kC}, $q^{\dagger}_{kC}$, $\bar{q}_{kC}$ and $\bar{q}^{\dagger}_{kC}$ are the prescribed Neumann mass fluxes of species k on $\mathit{\Gamma}_5 \subset \mathit{\Gamma}$. If $q_{kC} = 0$ and $\bar{q}_{kC} = 0$ the Neumann-type BC reduces to a *natural* (no-mass flux) BC associated with a zero concentration gradient $\nabla C_k = \mathbf{0}$ for the convective form of the mass transport equation, sometimes called as *Danckwerts condition* [111]. Alternatively, however, for the divergence form of the mass transport equation, if $q^{\dagger}_{kC} = 0$ and $\bar{q}^{\dagger}_{kC} = 0$ the Neumann-type BC reduces to a *natural* (no-mass flux) BC which forces the total (advective plus dispersive) mass flux to zero on $\varGamma_5$. Both variants of BC have their advantages. While the Neumann-type BC for the convective form is easier to implement and more flexible, their counterparts for the divergence form provide a stronger formulation in terms of mass convervation, however, possess difficulties at outflow boundaries on which the total mass flux is unknown (see Sect. 6.5.7).

The divergence form is capable of prescribing the total mass flux along a boundary portion resulting from the advective (convective) part $C_k\, q_{n_h}$ (load of concentration C_k in the liquid flow $q_{n_h} = \boldsymbol{q} \cdot \boldsymbol{n}$) and the dispersive part $-(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n}$. However, regarding this formulation all boundaries have to be specified with such type of BC, which can cause a specific handling of such formulations in the case of unknown mass concentration C_k on outflow boundaries (rather, C_k is here to be solved). Such boundaries require a specific treatment. This is done by evaluating the liquid flux via a budget analysis in a postprocessing step of computation which is then involved in modifying the BC of the mass flux at such portions of boundaries, for more see discussion in Sect. 6.5.7.

On the other hand, the default convective form does not require a specific handling associated with formulations on outflow boundaries and is usually preferred. Assigning $q_{kC} = -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} \approx 0$ as a natural BC, the mass flux freely passes through an advectively open boundary section and the concentration on the boundary automatically results. Note here, a boundary source of mass, as far as it should not be modeled via a 1st kind BC, in form of a mass boundary flux $q_{kC} \neq 0$ includes only the dispersive part, i.e., the magnitude of the flux will result from the gradient of concentration at the boundary. Thus in general, the convective form will necessarily produce a higher concentration gradient to realize the same mass load through a boundary.

However, there is a way to formulate mass flux BC providing an advective load of mass in form of Cauchy-type BC even for the convective form of mass transport. Indeed, we need not to resort to the divergence form in order to achieve suited mass load conditions on boundaries. It is easy to see that the Neumann-type BC for the divergence form, e.g., (6.19), is equivalent to Cauchy-type BC written as

$$\begin{aligned} -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= q_{kC}^{\dagger} - C_k\, q_{n_h} \\ &= q_{n_h} (C_{kC} - C_k) \end{aligned} \tag{6.21}$$

with known q_{n_h} and $q_{kC}^{\dagger} \approx q_{n_h} C_{kC}$ approximated as an input advective mass flux with prescribed boundary concentration C_{kC} for the convective form as further discussed in Sect. 6.3.2.3.

6.3.2.3 ⊗ Cauchy-Type and Robin-Type (3rd Kind) BC

For 3D and 2D (vertical and axisymmetric):

$$\left.\begin{aligned}&\text{convective form}\\ &q_{n_{kC}}(\boldsymbol{x},t) = -(\boldsymbol{D}_k\cdot\nabla C_k)\cdot\boldsymbol{n} \qquad\qquad = -\Phi_{kC}(C_{kC}-C_k)\\ &\text{divergence form}\\ &q_{n_{kC}}(\boldsymbol{x},t) = C_k\, q_{n_h} - (\boldsymbol{D}_k\cdot\nabla C_k)\cdot\boldsymbol{n} = -\Phi^{\dagger}_{kC}(C_{kC}-C_k)\end{aligned}\right\} \quad \text{on} \quad \Gamma_6\times t[t_0,\infty) \tag{6.22}$$

and for 2D horizontal (confined and unconfined):

$$\left.\begin{aligned}&\text{convective form}\\ &\bar{q}_{n_{kC}}(\boldsymbol{x},t) = -(\bar{\boldsymbol{D}}_k\cdot\nabla C_k)\cdot\boldsymbol{n} \qquad\qquad = -\bar{\Phi}_{kC}(C_{kC}-C_k)\\ &\text{divergence form}\\ &\bar{q}_{n_{kC}}(\boldsymbol{x},t) = C_k\, \bar{q}_{n_h} - (\bar{\boldsymbol{D}}_k\cdot\nabla C_k)\cdot\boldsymbol{n} = -\bar{\Phi}^{\dagger}_{kC}(C_{kC}-C_k)\end{aligned}\right\} \quad \text{on} \quad \Gamma_6\times t[t_0,\infty) \tag{6.23}$$

where C_{kC} are prescribed values of species k concentration on $\Gamma_6 \subset \Gamma$. The signs of $q_{n_{kC}}$ and $\bar{q}_{n_{kC}}$ are chosen that the boundary mass fluxes are positive outward-directed if $C_k > C_{kC}$. In (6.22) and (6.23) the *mass transfer coefficients* Φ_{kC}, $\bar{\Phi}_{kC}$, $\Phi^{\dagger}_{kC}$ and $\bar{\Phi}^{\dagger}_{kC}$ represent *dual* directional functions in form of:

$$\Phi_{kC} = \begin{cases}\Phi^{\text{in}}_{kC}(\boldsymbol{x},t) & \text{for} \quad C_{kC} > C_k\\ \Phi^{\text{out}}_{kC}(\boldsymbol{x},t) & \text{for} \quad C_{kC} \le C_k\end{cases} \tag{6.24}$$

$$\bar{\Phi}_{kC} = \begin{cases}\bar{\Phi}^{\text{in}}_{kC}(\boldsymbol{x},t) & \text{for} \quad C_{kC} > C_k\\ \bar{\Phi}^{\text{out}}_{kC}(\boldsymbol{x},t) & \text{for} \quad C_{kC} \le C_k\end{cases} \tag{6.25}$$

and similar for $\Phi^{\dagger}_{kC}$ and $\bar{\Phi}^{\dagger}_{kC}$, which are in general functions of space $\boldsymbol{x}$ and time t. Accordingly, in specifying two alternate (if necessary temporal) transfer coefficients different transfer conditions can be input to distinguish between inflow conditions ($q_{n_{kC}} < 0$) and outflow conditions ($q_{n_{kC}} > 0$). The special case, e.g., $\Phi_{kC} = \Phi^{\text{in}}_{kC} = \Phi^{\text{out}}_{kC}$ (and similar for $\bar{\Phi}_{kC}$, $\Phi^{\dagger}_{kC}$ and $\bar{\Phi}^{\dagger}_{kC}$) does not differ between inward and outward mass boundary flux.

As already discussed in Sect. 2.2.2 the 3rd kind BC of the convective forms of (6.22) and (6.23) can be identified as Cauchy-type BC, while the 3rd kind BC of the divergence form represents a Robin-type (mixed) BC, which is most general. It has been shown by (6.21) that Neumann-type BC of the divergence form is equivalent to Cauchy-type BC of the convective form if we simply set

$$\Phi_{kC} = -q_{n_h} \tag{6.26}$$

where $q_{n_h} = \boldsymbol{q}\cdot\boldsymbol{n}$ is a known (positive outward directed) flux of liquid on Γ_6. A typical application of such type of BC is a leaky deposit, from where a mass flux intrudes into an aquifer with a given (advective) rate as schematized in Fig. 6.5. It is assumed that the deposit having a known concentration C_{kC} leaks by a given *load* and intrudes into the domain Ω through Γ_6 via

$$q^{\text{load}}_{kC} = q^{\text{out}}_h\, C_{kC} \tag{6.27}$$

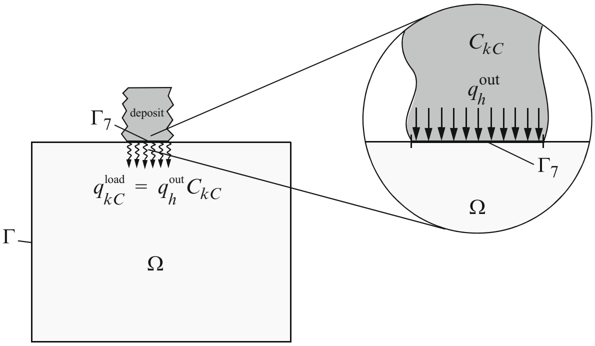


Fig. 6.5 Leak of a deposit: BC formulation of species mass load q_{kC}^{load} on $\Gamma_6 \subset \Gamma$

where q_{kC}^{load} is the load of species k on Γ_6 and q_h^{out} is the inward-directed flux of liquid leaving the deposit with concentration C_{kC}. Since $q_h^{\text{out}} = -q_{n_h}$ (negative due to the inward direction on Γ_6) we obtain with (6.26), i.e., $\Phi_{kC} = q_h^{\text{out}}$, the following Cauchy-type BC for the load of mass

$$q_{n_{kC}}(\boldsymbol{x}, t) = -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} = -q_h^{\text{out}}(C_{kC} - C_k) \quad \text{on} \quad \Gamma_6 \times t[t_0, \infty) \tag{6.28}$$

applied to the convective form of mass transport.

The transfer coefficients, e.g., Φ_{kC}, associated with BC's of 3rd kind (6.22) can be regarded as *leaching* parameters which constrain the mass flux through the boundary. If $\Phi_{kC} = 0$ the boundary becomes impervious. On the other hand, using a very large value $\Phi_{kC} \to \infty$ the BC of 3rd kind is reduced to a Dirichlet-type (1st kind) BC with $C_k = C_{kC}$ on Γ_6. Such a leaching process is displayed in Fig. 6.6 for the example of a flow over a salt dome modeled with a diffusive input condition ($C_{kC} > C_k$). Considering a thickness d for the leaching body and applying the Fick's law (4.67) written in 1D in form of

$$q_{n_{kC}} \approx -D_{ko}^{\text{in}} \frac{\Delta C_k}{\Delta s} = -D_{ko}^{\text{in}} \frac{C_{kC} - C_k}{d} \tag{6.29}$$

the mass transfer coefficient Φ_{kC}^{in} can be assessed as

$$\Phi_{kC}^{\text{in}} = \frac{D_{ko}^{\text{in}}}{d} \tag{6.30}$$

and analogously to a horizontal problem as

$$\bar{\Phi}_{kC}^{\text{in}} = B\Phi_{kC}^{\text{in}} = B\frac{D_{ko}^{\text{in}}}{d} \tag{6.31}$$

Analogous assessments for Φ_{kC}^{out} and $\bar{\Phi}_{kC}^{\text{out}}$ result if the transition resistance differs between inflow (leaching) and outflow (releasing) conditions: $\Phi_{kC}^{\text{in}} \neq \Phi_{kC}^{\text{out}}$ ($\bar{\Phi}_{kC}^{\text{in}} \neq \bar{\Phi}_{kC}^{\text{out}}$).

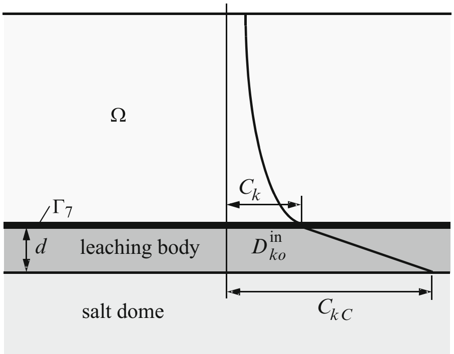


Fig. 6.6 Transfer coefficient Φ_{kC} $(= \Phi_{kC}^{\text{in}})$ as leaching parameter of a salt dome

6.3.2.4 ↱ Well-Type SPC

$$Q_{kw}(\boldsymbol{x},t) = -\sum_{w=1}^{N_\text{W}} C_{kw} Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) \quad \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0,\infty) \tag{6.32}$$

and

$$\int_\Omega Q_{kw} d\Omega = -\sum_{w=1}^{N_\text{W}} C_{kw} Q_w(t) \tag{6.33}$$

where Q_{kw} is the specific kth-species mass sink/source function of wells, $Q_w(t)$ is the prescribed volume per unit time discharge (pumping rate) of single well w pumped with a known concentration of C_{kw} at location $\boldsymbol{x}_w$ and $\delta(\boldsymbol{x}-\boldsymbol{x}_w) = \prod_{i=1}^{D} \delta(x_i - x_{iw})$ is the Dirac delta function associated with location $\boldsymbol{x}_w$. The well function Q_{kw} is assigned to a point sink of mass for the divergence form of mass transport equation.

In contrast, the convective form of mass transport has to be related to a well-point sink function in the following form (cf. mass transport equations of Table 3.7):

$$\begin{aligned} Q_{kw}(\boldsymbol{x},t) &= -\textstyle\sum_{w=1}^{N_\text{W}} C_{kw}(\boldsymbol{x}_w) Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) + C_k \ \textstyle\sum_{w=1}^{N_\text{W}} Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) \\ &= -\textstyle\sum_{w=1}^{N_\text{W}} Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w)\big(C_{kw} - C_k(\boldsymbol{x}_w)\big) \end{aligned} \tag{6.34}$$

and

$$\int_\Omega Q_{kw} d\Omega = -\sum_{w=1}^{N_\text{W}} Q_w(t)\big(C_{kw} - C_k(\boldsymbol{x}_w)\big) \tag{6.35}$$

which reveals a similarity to a Cauchy-like, however, point-related mass transfer relation as described above. Note that the pumping rate Q_w is positive for a sink (pump) and negative for a source (recharge/injection) at well point $\boldsymbol{x}_w$. These types of SPC in form of (6.32) and (6.34) are usually applied to cases, where a mass flux given by a flow rate of $Q_w < 0$ and known concentration C_{kw} is injected through wells w.

6.3.3 *Heat Transport BC*

6.3.3.1 ○ Dirichlet-Type (1st Kind) BC

$$T(\boldsymbol{x},t) = T_D(t) \quad \text{on} \quad \Gamma_7 \times t[t_0,\infty) \tag{6.36}$$

where T_D are prescribed values of temperature on $\Gamma_7 \subset \Gamma$. For steady-state heat transport problems Dirichlet-type BC's (6.36) are usually required, i.e., $\Gamma_7 \neq \emptyset$, unless Cauchy-type BC's occur.

6.3.3.2 × Neumann-Type (2nd Kind) BC

For 3D and 2D (vertical and axisymmetric):

$$\left.\begin{aligned} &\text{convective form} \\ &q_{n_T}(\boldsymbol{x},t) = \underbrace{-(\boldsymbol{\Lambda}\cdot\nabla T)\cdot\boldsymbol{n}}_{\text{conductive flux}} = q_T(t) \\ &\text{divergence form} \\ &q_{n_T}(\boldsymbol{x},t) = \underbrace{\rho c(T-T_0)\,q_{n_h} - (\boldsymbol{\Lambda}\cdot\nabla T)\cdot\boldsymbol{n}}_{\text{total flux}} = q_T^{\dagger}(t) \end{aligned}\right\} \quad \text{on} \quad \Gamma_8 \times t[t_0,\infty) \tag{6.37}$$

and for 2D horizontal (confined and unconfined):

$$\left.\begin{aligned} &\text{convective form} \\ &\bar{q}_{n_T}(\boldsymbol{x},t) = \underbrace{-(\bar{\boldsymbol{\Lambda}}\cdot\nabla T)\cdot\boldsymbol{n}}_{\text{conductive flux}} = \bar{q}_T(t) \\ &\text{divergence form} \\ &\bar{q}_{n_T}(\boldsymbol{x},t) = \underbrace{\rho c(T-T_0)\,\bar{q}_{n_h} - (\bar{\boldsymbol{\Lambda}}\cdot\nabla T)\cdot\boldsymbol{n}}_{\text{total flux}} = \bar{q}_T^{\dagger}(t) \end{aligned}\right\} \quad \text{on} \quad \Gamma_8 \times t[t_0,\infty) \tag{6.38}$$

where q_{n_T} and $\bar{q}_{n_T}$ represent normal heat fluxes (positive outward-directed) across the boundary Γ_8, T_0 is a reference temperature and q_T, $q_T^{\dagger}$, $\bar{q}_T$ and $\bar{q}_T^{\dagger}$ are the prescribed Neumann heat fluxes on $\varGamma_8 \subset \varGamma$. If $q_T = 0$ and $\bar{q}_T = 0$ the Neumann-type BC reduces to a *natural* (no-heat flux) *adiabatic BC* associated with a zero temperature gradient $\nabla T = \mathbf{0}$ for the convective form of the heat transport equation. Alternatively, however, for the divergence form of the heat transport equation, if $q_T^\dagger = 0$ and $\bar{q}_T^\dagger = 0$ the Neumann-type BC reduces to a *natural* (no-heat flux) BC which forces the total (advective plus conductive) heat flux to zero on $\varGamma_8$. The advantages of both variants of Neumann-type BC are already discussed in Sect. 6.3.2.2 in the context of mass transport. Similarly, the equivalence of the Neumann-type BC for the divergence form to the Cauchy-type BC for the convective form of the heat transport equation leads to the formulation of a heat load condition

$$\begin{aligned} -(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} &= q_T^\dagger - \rho c (T - T_0)\, q_{n_h} \\ &= \rho c q_{n_h} (T_C - T) \end{aligned} \tag{6.39}$$

with known $q_{n_h} = \boldsymbol{q} \cdot \boldsymbol{n}$ and $q_T^\dagger \approx \rho c q_{n_h} (T_C - T_0)$ approximated as an input advective heat flux with prescribed boundary temperature difference $T_C - T_0$ for the convective form as further discussed in Sect. 6.3.3.3.

6.3.3.3 ⊗ Cauchy-Type and Robin-Type (3rd Kind) BC

For 3D and 2D (vertical and axisymmetric):

$$\left.\begin{aligned} &\text{convective form} \\ &q_{n_T}(\boldsymbol{x}, t) = -(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} \qquad\qquad\qquad = -\varPhi_T (T_C - T) \\ &\text{divergence form} \\ &q_{n_T}(\boldsymbol{x}, t) = \rho c (T - T_0)\, q_{n_h} - (\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} = -\varPhi_T^\dagger (T_C - T) \end{aligned}\right\} \quad \text{on} \quad \varGamma_9 \times t[t_0, \infty) \tag{6.40}$$

and for 2D horizontal (confined and unconfined):

$$\left.\begin{aligned} &\text{convective form} \\ &\bar{q}_{n_T}(\boldsymbol{x}, t) = -(\bar{\boldsymbol{\Lambda}} \cdot \nabla T) \cdot \boldsymbol{n} \qquad\qquad\qquad = -\bar{\varPhi}_T (T_C - T) \\ &\text{divergence form} \\ &\bar{q}_{n_T}(\boldsymbol{x}, t) = \rho c (T - T_0)\, \bar{q}_{n_h} - (\bar{\boldsymbol{\Lambda}} \cdot \nabla T) \cdot \boldsymbol{n} = -\bar{\varPhi}_T^\dagger (T_C - T) \end{aligned}\right\} \quad \text{on} \quad \varGamma_9 \times t[t_0, \infty) \tag{6.41}$$

where T_C are prescribed values of temperature on $\varGamma_9 \subset \varGamma$. The signs of q_{n_T} and $\bar{q}_{n_T}$ are chosen that the boundary heat fluxes are positive outward-directed if $T > T_C$. In (6.40) and (6.41) the *heat transfer coefficients* $\varPhi_T$, $\bar{\varPhi}_T$, $\varPhi_T^\dagger$ and $\bar{\varPhi}_T^\dagger$ represent *dual* directional functions in form of:

$$\Phi_T = \begin{cases} \Phi_T^{\text{in}}(\boldsymbol{x}, t) & \text{for} \quad T_C > T \\ \Phi_T^{\text{out}}(\boldsymbol{x}, t) & \text{for} \quad T_C \leq T \end{cases} \tag{6.42}$$

$$\bar{\Phi}_T = \begin{cases} \bar{\Phi}_T^{\text{in}}(\boldsymbol{x}, t) & \text{for} \quad T_C > T \\ \bar{\Phi}_T^{\text{out}}(\boldsymbol{x}, t) & \text{for} \quad T_C \leq T \end{cases} \tag{6.43}$$

and similar for $\Phi_T^\dagger$ and $\bar{\Phi}_T^\dagger$, which are in general functions of space $\boldsymbol{x}$ and time t. Accordingly, in specifying two alternate (if necessary temporal) transfer coefficients different transfer conditions can be input to distinguish between inflow conditions ($q_{n_T} < 0$) and outflow conditions ($q_{n_T} > 0$). The special case, e.g., $\Phi_T = \Phi_T^{\text{in}} = \Phi_T^{\text{out}}$ (and similar for $\bar{\Phi}_T$, $\Phi_T^\dagger$ and $\bar{\Phi}_T^\dagger$) does not differ between inward and outward heat boundary flux.

The 3rd kind BC of the convective forms of (6.40) and (6.41) can be identified as Cauchy-type BC, while the 3rd kind BC of the divergence form represents a Robin-type (mixed) BC, which is most general (cf. Sect. 2.2.2). It has been shown by (6.39) that Neumann-type BC of the divergence form is equivalent to Cauchy-type BC of the convective form if we simply set $\Phi_T = -\rho c q_{n_h}$, where $q_{n_h} = \boldsymbol{q} \cdot \boldsymbol{n}$ is a known (positive outward directed) flux of liquid on Γ_9. This allows to prescribe (similar to the mass transport in Sect. 6.3.2.3) a heat load BC, viz.,

$$q_{n_T}(\boldsymbol{x}, t) = -(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} = -\rho c q_h^{\text{out}} (T_C - T) \quad \text{on} \quad \Gamma_9 \times t[t_0, \infty) \tag{6.44}$$

applied to the convective form of heat transport, where the heat load $q_T^{\text{load}} = q_h^{\text{out}} \rho c (T_C - T_0)$ on Γ_9 is forced by the inward-directed flux of liquid $q_h^{\text{out}} = -q_{n_h}$ entering with a boundary temperature T_C.

The heat transfer coefficients, e.g., Φ_T, associated with BC's of 3rd kind (6.40) represent *heat transition* parameters. If $\Phi_T = 0$ the boundary becomes adiabatic (insulated). On the other hand, using a very large value $\Phi_T \to \infty$ the BC of 3rd kind is reduced to a Dirichlet-type (1st kind) BC with $T = T_C$ on Γ_9. The heat transfer coefficients can be estimated analogously to the above transfer coefficients for mass flux of Sect. 6.3.2.3. Considering a thickness d for a heat transition layer and applying Fourier's law (4.76) for input condition ($T_C > T$) in form of:

$$q_{n_T} \approx -\Lambda_o^{\text{in}} \frac{\Delta T}{\Delta s} = -\Lambda_o^{\text{in}} \frac{T_C - T}{d} \tag{6.45}$$

the heat transfer coefficient Φ_T^{in} can be obtained as

$$\Phi_T^{\text{in}} = \frac{\Lambda_o^{\text{in}}}{d} \tag{6.46}$$

and similarly to a horizontal problem as

$$\bar{\Phi}_T^{\text{in}} = B \Phi_T^{\text{in}} = B \frac{\Lambda_o^{\text{in}}}{d} \tag{6.47}$$

where Λ_o^{in} represents the heat conduction coefficient of the transition layer. Analogous assessments for Φ_T^{out} and $\bar{\Phi}_T^{\mathrm{out}}$ result if the heat transition resistance differs between inflow (leaching) and outflow (releasing) conditions: $\Phi_T^{\mathrm{in}} \neq \Phi_T^{\mathrm{out}}$ ($\bar{\Phi}_T^{\mathrm{in}} \neq \bar{\Phi}_T^{\mathrm{out}}$).

More general heat transfer coefficients and related thermal resistances of transition layers are described in Appendix E for single and composite plane wall and circular pipe wall configurations. It results in heat transfer coefficients exemplified in the form

$$\Phi_T = \frac{1}{S \sum_i R_i} \tag{6.48}$$

with the *specific thermal resistance* R_i of solid material i given as

$$R_i = \begin{cases} \dfrac{d_i}{S\,\Lambda_i^s} & \text{plane wall} \\ \dfrac{\ln(r_{i+1}/r_i)}{2\pi\ \Lambda_i^s} & \text{circular pipe wall} \end{cases} \tag{6.49}$$

where S is the specific exchange area and Λ_i^s is the thermal conductivity of solid material i. Note that for pipe wall geometry $S = 2\pi r$, where r is the radius of the boundary surface at Γ_9.

6.3.3.4 ↱ Well-Type SPC

$$Q_{Tw}(\boldsymbol{x},t) = -\sum_{w=1}^{N_\mathrm{W}} (T_w - T_0)\,\rho c\, Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) \quad \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0,\infty) \tag{6.50}$$

and

$$\int_\Omega Q_{Tw} d\Omega = -\sum_{w=1}^{N_\mathrm{W}} (T_w - T_0)\,\rho c\, Q_w(t) \tag{6.51}$$

where Q_{Tw} is the specific heat sink/source function of wells, $Q_w(t)$ is the prescribed volume per unit time discharge (pumping rate) of single well w pumped with a known temperature of T_w at location $\boldsymbol{x}_w$, $\delta(\boldsymbol{x}-\boldsymbol{x}_w) = \prod_{i=1}^{D} \delta(x_i - x_{iw})$ is the Dirac delta function associated with location $\boldsymbol{x}_w$ and T_0 is the reference temperature. The well function Q_{Tw} is assigned to a point sink of heat for the divergence form of heat transport equation.

In contrast, the convective form of heat transport has to be related to a well-point sink function in the following form (cf. heat transport equations of Table 3.7):

$$
\begin{aligned}
Q_{Tw}(\boldsymbol{x},t) &= -\textstyle\sum_{w=1}^{N_{\mathrm{W}}}\big(T_w(\boldsymbol{x}_w)-T_0\big)\rho c\, Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w)+ \\
&\qquad \rho c(T-T_0)\ \textstyle\sum_{w=1}^{N_{\mathrm{W}}} Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) \\
&= -\textstyle\sum_{w=1}^{N_{\mathrm{W}}} \rho c\, Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w)\big(T_w - T(\boldsymbol{x}_w)\big)
\end{aligned}
\tag{6.52}
$$

and

$$
\int_{\Omega} Q_{Tw}\, d\Omega = -\sum_{w=1}^{N_{\mathrm{W}}} \rho c\, Q_w(t)\big(T_w - T(\boldsymbol{x}_w)\big) \tag{6.53}
$$

which reveals a similarity to a Cauchy-like, however, point-related heat transfer relation as described above. Note that the pumping rate Q_w is positive for a sink (pump) and negative for a source (recharge/injection) at well point $\boldsymbol{x}_w$. These types of SPC in form of (6.50) and (6.52) are usually applied to cases, where a heat flux given by a flow rate of $Q_w < 0$ and known temperature T_w is injected through wells w.

6.4 BC Constraints (BCC's) and SPC Constraints (SPCC's)

Constraints are limitations for all types of BC's and SPC's. They can be written for BC's in the following form:

$$
\text{value of BC is valid } \textit{if} \begin{cases} < \text{Max bound(s) } \textit{else} \text{ replace BC by Max bound} \\ \quad \text{and} \\ > \text{Min bound(s) } \textit{else} \text{ replace BC by Min bound} \end{cases} \tag{6.54}
$$

They result from the requirement that BC should only be valid as long as minimum and maximum bounds are satisfied. If during a simulation run the conditions are violated, the constraints are to be assigned as new intermediate BC. The same procedure is applied to SPC's.

The formulation of constraints is commonly based on the formalism of *complementary conditions* for a type of BC and SPC. Accordingly, value-type (1st kind and 3rd kind) BC's (hydraulic head, species concentration or temperature) are constrained by maximum and minimum flux relations (liquid, mass and heat fluxes, respectively). On the other hand, flux-type (2nd kind) BC's and well-type SPC's are constrained by complementary limits of boundary values, i.e., the liquid flux is constrained by maximum-minimum hydraulic heads, the mass flux by minimum-maximum species concentrations and the heat flux by minimum-maximum temperatures. Following formulations are available for flow, mass and heat conditions.

6.4.1 Flow BCC and SPCC

$$
\begin{array}{llll}
\overline{\underline{\bigcirc}} & \text{1st kind} & h_D(t) & \textit{if} \begin{cases} \text{Max: } Q_{n_h} < Q_{n_h}^{\max_1}(t) \ \textit{else} \ Q_{n_h} = Q_{n_h}^{\max_1}(t) \\ \text{Min: } Q_{n_h} > Q_{n_h}^{\min_1}(t) \ \textit{else} \ Q_{n_h} = Q_{n_h}^{\min_1}(t) \end{cases} \\
\overline{\underline{\times}} & \text{2nd kind} & q_h(t) & \textit{if} \begin{cases} \text{Max: } h < h^{\max_2}(t) \quad \textit{else} \ h = h^{\max_2}(t) \\ \text{Min: } h > h^{\min_2}(t) \quad \textit{else} \ h = h^{\min_2}(t) \end{cases} \\
\overline{\underline{\otimes}} & \text{3rd kind} & h_C(t) & \textit{if} \begin{cases} \text{Max: } Q_{n_h} < Q_{n_h}^{\max_3}(t) \ \textit{else} \ Q_{n_h} = Q_{n_h}^{\max_3}(t) \\ \text{Min: } Q_{n_h} > Q_{n_h}^{\min_3}(t) \ \textit{else} \ Q_{n_h} = Q_{n_h}^{\min_3}(t) \end{cases} \\
\overline{\underline{\vdash}} & \text{well type} & Q_{hw}(t) & \textit{if} \begin{cases} \text{Max: } h < h^{\max_4}(t) \quad \textit{else} \ h = h^{\max_4}(t) \\ \text{Min: } h > h^{\min_4}(t) \quad \textit{else} \ h = h^{\min_4}(t) \end{cases}
\end{array}
\tag{6.55}
$$

where

$$
Q_{n_h} = -\int q_{n_h} d\Gamma \tag{6.56}
$$

represents the integral boundary balance flux of liquid summed-up at discrete (nodal) points to which the corresponding boundary values are related. Note, due to compatibility reasons with SPC's the pointwise balance quantity is defined negative outward (because a positive SPC acts as a sink). The flux Q_{n_h} has to be computed in a balance analysis during the simulation (cf. Sect. 8.19.2). The minimum-maximum bounds $Q_{n_h}^{\min_1}$, $Q_{n_h}^{\max_1}$, $h^{\min_2}$, $h^{\max_2}$, $Q_{n_h}^{\min_3}$, $Q_{n_h}^{\max_3}$, $h^{\min_4}$ and $h^{\max_4}$ are optional input parameters and can be even time-dependent functions. Accordingly, it is possible to consider time-dependent variations in the existence and influence of boundary and constraint conditions. For instance, these temporary capabilities of constraints are very useful in modeling the temporarily varying occurrence of sealing or drainage activities over a restricted time period, or in simulating time-constrained BC's (e.g., 1st kind) which are associated with certain construction or remedial actions arising only at given times. Typical applications of constraint conditions formulated by (6.55) are shown in sketches of Fig. 6.7.

In the first example (Fig. 6.7a) a single well operation is constrained by minimum and maximum head conditions. A well-type SPC with a given recharging or extracting discharge Q_{hw} is applied. The computation results a hydraulic head h at the borehole. Only if the resulting head is between the bounds $h^{\min_4}$ and $h^{\max_4}$ the computation is accepted, otherwise if the head h is smaller than $h^{\min_4}$ the SPC is replaced by $h = h^{\min_4}$ at the point, which represents a (pointwise) Dirichlet-type BC, and the computation has to be repeated for the changed BC. Similarly, if the resulting head h is larger than $h^{\max_4}$ the SPC is replaced by the $h = h^{\max_4}$ Dirichlet-type BC at the point and the solution has to be restarted again.

The second example (Fig. 6.7b) is regarded to a flux-limited infiltration through a river bed. A 3rd kind BC with a hydraulic head h_C of the river is applied and constrained by a maximum flux $Q_{n_h}^{\max_3}$. If the groundwater table decreases below the

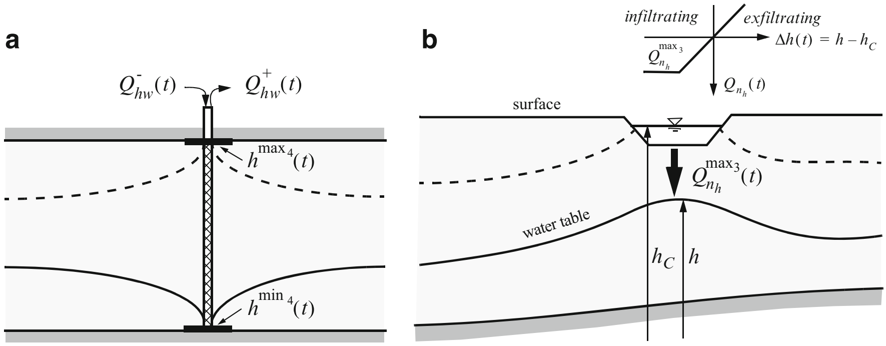


Fig. 6.7 Examples of using constraints for flow problems: (**a**) constraining a single well by an allowable drawdown in form of a minimum well head and by an allowable injection water level in form of a maximum well head, (**b**) flow separation in infiltration from surface water due to constraining the maximum seepage through the river bed

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 of a flow transfer in form of (6.7) for the infiltrating water as a function of the difference $\Delta h = h - h_C$ between the groundwater head h and the reference (river) head h_C cannot be maintained anymore. It requires the prescription of the maximum bound $Q^{\max_3}_{n_h}$. The formulation is termed as *flux-constrained transfer* BC. In this case the computation is started with the given 3rd kind BC. After the computation balance fluxes Q_{n_h} at the boundary are evaluated. If Q_{n_h} violates the maximum bound $Q^{\max_3}_{n_h}$ (or the minimum bound $Q^{\min_3}_{n_h}$) the computation has to be repeated with changed BC in form of $Q_{hw} = Q_{n_h} = Q^{\max_3}_{n_h}$ (or $Q_{hw} = Q_{n_h} = Q^{\min_3}_{n_h}$), which represent a well-type SPC.

Although flux-constrained transfer BC's are quite general formulations, their specification is sometimes cumbersome because the determination of the constraint fluxes requires geometric information of the boundaries (e.g., transfer areas). A more convenient and alternative formulation of constraints for 3rd kind BC's is in the form of the so-called *head-constrained transfer* BC as exemplified in Fig. 6.8 for a flux-limiting infiltration through a river bed. Instead of a direct input of constraint fluxes according to (6.55), maximum and minimum head values $h^{\max}_C$ and $h^{\min}_C$, respectively, are prescribed, which are used to derive the constrained min-max fluxes for Cauchy-type BC's. It reads as follows:

$$\underline{\overline{\otimes}} \quad \text{3rd kind} \;\; h_C(t) \;\; \mathit{if} \;\; \begin{cases} \text{Max:} \;\; h < h^{\max}_C(t) \;\; \mathit{else} \\ \qquad q_{n_h} = q^{\min}_h = -\Phi_h (h_C - h^{\max}_C) \;\; \mathit{if} \;\; h_C \leq h^{\max}_C \\ \\ \text{Min:} \;\; h > h^{\min}_C(t) \;\; \mathit{else} \\ \qquad q_{n_h} = q^{\max}_h = -\Phi_h (h_C - h^{\min}_C) \;\; \mathit{if} \;\; h_C \geq h^{\min}_C \end{cases} \tag{6.57}$$

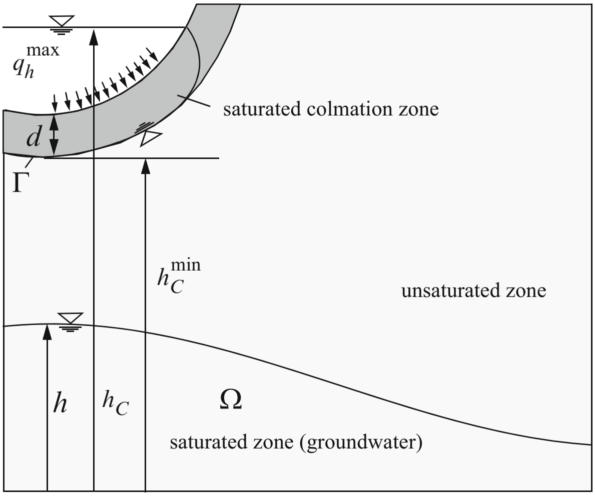


Fig. 6.8 Head-constrained transfer BC for a flux-limiting infiltration through a river bed

Note that the effects of the constraints in (6.55) and (6.57) are different. It is apparent that the minimum head bound $h_C^{\min}$ determines the maximum flux rate $q_h^{\max}$ and the maximum head bound $h_C^{\max}$ yields the minimum flux rate $q_h^{\min}$.

The advantage of head-based constraint formulation is that the limiting (constraint) fluxes are rates and no more integral balance fluxes, which makes the computation more efficient. The transfer coefficient Φ_h in (6.57) can be determined from the layer parameters of the clogged river bed as discussed in Sect. 6.3.1.2. 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 $h_C = h_C(t)$ a temporal no flux condition is automatically satisfied if the reference head h_C becomes identical to the constrained head $h_C^{\min}$ (or $h_C^{\max}$) in time. It means written for the minimum constraint

$$\begin{gathered} q_{n_h} = q_h^{\max} \equiv 0 \quad \text{for} \\ h_C(t) = h_C^{\min}(t) \quad \text{and} \quad h(t) < h_C^{\min} \end{gathered} \tag{6.58}$$

To force a temporal no flux condition independent of the groundwater head $h = h(t)$, the maximum head constraint has to be set additionally to the reference head. It requires

$$\begin{gathered} q_{n_h} = q_h^{\max} \equiv 0 \quad \text{for} \\ h_C(t) = h_C^{\min}(t) = h_C^{\max}(t) \quad \text{and arbitrary} \quad h(t) \end{gathered} \tag{6.59}$$

6.4.2 Mass Transport BCC and SPCC

$$\begin{array}{llll}
\underline{\overline{\bigcirc}}\ \text{1st kind} & C_{kD}(t) & \mathit{if} & \begin{cases} \text{Max:} \\ \left\{\begin{array}{c} Q_{n_kC}<Q_{n_kC}^{\max_1}(t) \\ \mathit{and} \\ h^{\min_1}\le h\le h^{\max_1}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} Q_{n_kC}=Q_{n_kC}^{\max_1}(t)\\ \mathit{as\ long\ as}\ h^{\min_1}\le h\le h^{\max_1};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_1}\ \mathit{or}\ h>h^{\max_1}\end{array}\right\} \\ \text{Min:} \\ \left\{\begin{array}{c} Q_{n_kC}>Q_{n_kC}^{\min_1}(t) \\ \mathit{and} \\ h^{\min_1}\le h\le h^{\max_1}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} Q_{n_kC}=Q_{n_kC}^{\min_1}(t)\\ \mathit{as\ long\ as}\ h^{\min_1}\le h\le h^{\max_1};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_1}\ \mathit{or}\ h>h^{\max_1}\end{array}\right\} \end{cases} \\
\underline{\overline{\times}}\ \text{2nd kind} & q_{kC}(t) & \mathit{if} & \begin{cases} \text{Max:} \\ \left\{\begin{array}{c} C_k<C_k^{\max_2}(t) \\ \mathit{and} \\ h^{\min_2}\le h\le h^{\max_2}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} C_k=C_k^{\max_2}(t)\\ \mathit{as\ long\ as}\ h^{\min_2}\le h\le h^{\max_2};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_2}\ \mathit{or}\ h>h^{\max_2}\end{array}\right\} \\ \text{Min:} \\ \left\{\begin{array}{c} C_k>C_k^{\min_2}(t) \\ \mathit{and} \\ h^{\min_2}\le h\le h^{\max_2}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} C_k=C_k^{\min_2}(t)\\ \mathit{as\ long\ as}\ h^{\min_2}\le h\le h^{\max_2};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_2}\ \mathit{or}\ h>h^{\max_2}\end{array}\right\} \end{cases} \\
\underline{\overline{\otimes}}\ \text{3rd kind} & C_{kC}(t) & \mathit{if} & \begin{cases} \text{Max:} \\ \left\{\begin{array}{c} Q_{n_kC}<Q_{n_kC}^{\max_3}(t) \\ \mathit{and} \\ h^{\min_3}\le h\le h^{\max_3}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} Q_{n_kC}=Q_{n_kC}^{\max_3}(t)\\ \mathit{as\ long\ as}\ h^{\min_3}\le h\le h^{\max_3};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_3}\ \mathit{or}\ h>h^{\max_3}\end{array}\right\} \\ \text{Min:} \\ \left\{\begin{array}{c} Q_{n_kC}>Q_{n_kC}^{\min_3}(t) \\ \mathit{and} \\ h^{\min_3}\le h\le h^{\max_3}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} Q_{n_kC}=Q_{n_kC}^{\min_3}(t)\\ \mathit{as\ long\ as}\ h^{\min_3}\le h\le h^{\max_3};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_3}\ \mathit{or}\ h>h^{\max_3}\end{array}\right\} \end{cases} \\
\underline{\overline{\Rsh}}\ \text{well type} & Q_{kw}(t) & \mathit{if} & \begin{cases} \text{Max:} \\ \left\{\begin{array}{c} C_k<C_k^{\max_4}(t) \\ \mathit{and} \\ h^{\min_4}\le h\le h^{\max_4}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} C_k=C_k^{\max_4}(t)\\ \mathit{as\ long\ as}\ h^{\min_4}\le h\le h^{\max_4};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_4}\ \mathit{or}\ h>h^{\max_4}\end{array}\right\} \\ \text{Min:} \\ \left\{\begin{array}{c} C_k>C_k^{\min_4}(t) \\ \mathit{and} \\ h^{\min_4}\le h\le h^{\max_4}\end{array}\right\} \mathit{else} \left\{\begin{array}{c} C_k=C_k^{\min_4}(t)\\ \mathit{as\ long\ as}\ h^{\min_4}\le h\le h^{\max_4};\\ Q_{n_kC}=0\ \mathit{if}\ h<h^{\min_4}\ \mathit{or}\ h>h^{\max_4}\end{array}\right\} \end{cases}
\end{array} \tag{6.60}$$

where

$$Q_{n_kC} = -\int q_{n_kC}\, d\Gamma \tag{6.61}$$

represents the integral boundary balance mass flux of species k summed-up at discrete (nodal) points to which the corresponding boundary values are related (cf. Sect. 8.19.2), $(\ldots)^{\max_i}$ and $(\ldots)^{\min_i}$ denote the prescribed maximum and minimum bounds, respectively, for the corresponding type of BC and SPC, and C_k and h in (6.60) are the concentration of species k and the hydraulic head, respectively, computed on the boundary or the singular point. The min-max bounds for the flux

$Q_{n_{kC}}$, the concentration C_k and the hydraulic head h can be again time-dependent functions allowing very comfortable rules for constraints.

Naturally, the specific balance mass flux $q_{n_{kC}}$ used in (6.61) is composed of the advective and dispersive parts according to

$$q_{n_{kC}} = \underbrace{C_k\, q_{n_h}}_{\text{advective}} - \underbrace{(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n}}_{\text{dispersive}} \tag{6.62}$$

In practice, it has shown to be inappropriate to include the total (advective plus dispersive) flux into the procedure of controlling the constraint conditions (6.60) because the direction of the dispersive fluxes is ambiguous (e.g., the dispersive spreading also occurs against the advective flow direction). Accordingly, the balance-based evaluation of fluxes has to be exclusively related to the advective mass fluxes, viz.,

$$Q_{n_{kC}} = -\int q_{n_{kC}} d\Gamma \approx -\int (C_k\, q_{n_h}) d\Gamma \tag{6.63}$$

presenting unambiguously directional balance quantities.

The transport constraints (6.60) essentially consist of two parts for the individual types of BC's and SPC's:

1. A min-max bound complementary for the type of BC and SPC is imposed, i.e., a concentration boundary (1st or 3rd kind) is controlled by an allowable min-max boundary mass flux, and a mass flux boundary magnitude (2nd kind or well type) is controlled by an allowable min-max boundary concentration.
2. Optionally, a permitted range for BC and SPC within tolerable limits of hydraulic head h (ranging between $h^{\min_i}$ and $h^{\max_i}$) is imposed. If the simulated water table h lies outside this range, the BC's (all types, 1st to 3th kind) and SPC's are suppressed. This can easily be realized by assigning intermediately a zero flux $Q_{n_{kC}} = 0$, i.e., no mass flux then occurs and the BC's and SPC's are switched off.

Typical applications of mass transport constraints are outlined in Fig. 6.9. Figure 6.9a describes the case of a density-coupled saltwater intrusion problem (flow over a salt dome) having a boundary on which alternate boundary concentrations appear in dependence on the dynamic process: As long as water enters the domain it should have a prescribed concentration of freshwater. However, if the water releases the domain (along the same boundary) the concentration on the boundary is unknown and should be automatically computed. Such a description can be easily realized if the entire boundary section is assigned by a freshwater BC of 1st kind $C_k = C_{kD}$, and at the same time, the boundary will be imposed by a constraint condition in form of a null minimum mass flux $Q_{n_{kC}}^{\min_1} = 0$ (more constraints are not necessarily to be specified). Such an arrangement provides that the freshwater condition remains valid as long as the advective (convective) flux points into the

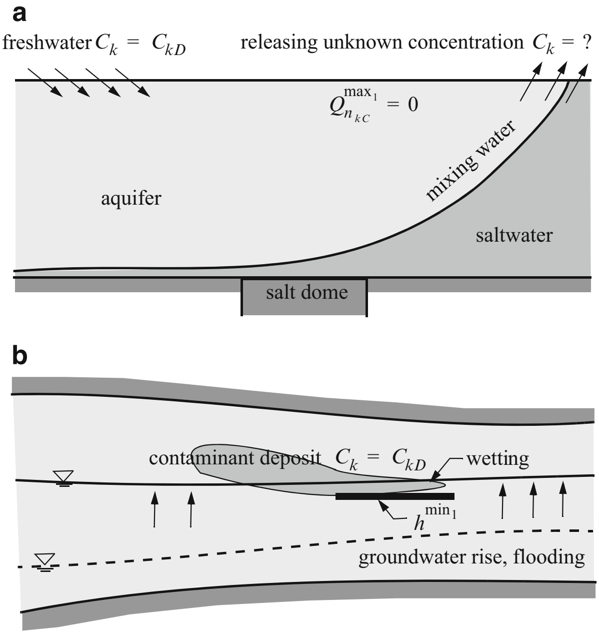

a
freshwater $C_k = C_{kD}$
releasing unknown concentration $C_k = ?$
$Q_{n_{kC}}^{\max_1} = 0$
mixing water
aquifer
saltwater
salt dome
b
contaminant deposit $C_k = C_{kD}$
wetting
$h^{\min_1}$
groundwater rise, flooding

Fig. 6.9 Application of mass transport constraints: (**a**) Saltwater intrusion by flowing groundwater over a salt dome and (**b**) wetting and activating a contaminant deposit during a groundwater rise (flooding)

domain[1]:

$$Q_{n_{kC}} = -\int (C_k\, q_{n_h}) d\Gamma \;>\; Q_{n_{kC}}^{\min_1} = 0 \qquad \text{since} \;\; q_{n_h} < 0 \;\; \text{for inflow} \tag{6.64}$$

The second example shown in Fig. 6.9b describes an application in modeling a contaminant spreading process from a deposit associated with rising groundwater in a phreatic aquifer (referred to as flooding problem). The contaminant BC (e.g., modeled as a 1st kind type) should be active only when the water table reaches the contaminant deposit (wetting case), i.e., a constraint in form of $h^{\min_1}$ is prescribed representing the bottom of the contaminant deposit. More constraints are not necessarily required in such a case.

[1] Note that a freshwater condition identical to zero ($C_{kD} = 0$) is inappropriate in the present balance-based computation to differ between inward and outward directed advective (convective) fluxes. It can fail because the directional magnitude of $Q_{n_{kC}}$ according to (6.64) is no more identifiable since $Q_{n_{kC}} = 0 \equiv Q_{n_{kC}}^{\min_1} = 0$! Accordingly, instead of zero it is recommended to use a numerically very small value for C_{kD}.

6.4.3 Heat Transport BCC and SPCC

$$
\begin{array}{llll}
\overline{\underline{\bigcirc}} & \text{1st kind} & T_D(t) & \mathit{if} \begin{cases}
\text{Max:}\\
\left\{\begin{matrix} Q_{n_T}<Q_{n_T}^{\max_1}(t)\\ \mathit{and}\\ h^{\min_1}\le h\le h^{\max_1}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} Q_{n_T}=Q_{n_T}^{\max_1}(t)\\ \mathit{as\ long\ as}\ h^{\min_1}\le h\le h^{\max_1};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_1}\ \mathit{or}\ h>h^{\max_1}\end{matrix}\right\}\\
\text{Min:}\\
\left\{\begin{matrix} Q_{n_T}>Q_{n_T}^{\min_1}(t)\\ \mathit{and}\\ h^{\min_1}\le h\le h^{\max_1}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} Q_{n_T}=Q_{n_T}^{\min_1}(t)\\ \mathit{as\ long\ as}\ h^{\min_1}\le h\le h^{\max_1};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_1}\ \mathit{or}\ h>h^{\max_1}\end{matrix}\right\}
\end{cases}\\
\overline{\underline{\times}} & \text{2nd kind} & q_T(t) & \mathit{if} \begin{cases}
\text{Max:}\\
\left\{\begin{matrix} T<T^{\max_2}(t)\\ \mathit{and}\\ h^{\min_2}\le h\le h^{\max_2}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} T=T^{\max_2}(t)\\ \mathit{as\ long\ as}\ h^{\min_2}\le h\le h^{\max_2};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_2}\ \mathit{or}\ h>h^{\max_2}\end{matrix}\right\}\\
\text{Min:}\\
\left\{\begin{matrix} T>T^{\min_2}(t)\\ \mathit{and}\\ h^{\min_2}\le h\le h^{\max_2}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} T=T^{\min_2}(t)\\ \mathit{as\ long\ as}\ h^{\min_2}\le h\le h^{\max_2};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_2}\ \mathit{or}\ h>h^{\max_2}\end{matrix}\right\}
\end{cases}\\
\overline{\underline{\otimes}} & \text{3rd kind} & T_C(t) & \mathit{if} \begin{cases}
\text{Max:}\\
\left\{\begin{matrix} Q_{n_T}<Q_{n_T}^{\max_3}(t)\\ \mathit{and}\\ h^{\min_3}\le h\le h^{\max_3}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} Q_{n_T}=Q_{n_T}^{\max_3}(t)\\ \mathit{as\ long\ as}\ h^{\min_3}\le h\le h^{\max_3};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_3}\ \mathit{or}\ h>h^{\max_3}\end{matrix}\right\}\\
\text{Min:}\\
\left\{\begin{matrix} Q_{n_T}>Q_{n_T}^{\min_3}(t)\\ \mathit{and}\\ h^{\min_3}\le h\le h^{\max_3}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} Q_{n_T}=Q_{n_T}^{\min_3}(t)\\ \mathit{as\ long\ as}\ h^{\min_3}\le h\le h^{\max_3};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_3}\ \mathit{or}\ h>h^{\max_3}\end{matrix}\right\}
\end{cases}\\
\overline{\underline{\ulcorner}} & \text{well type} & Q_{Tw}(t) & \mathit{if} \begin{cases}
\text{Max:}\\
\left\{\begin{matrix} T<T^{\max_4}(t)\\ \mathit{and}\\ h^{\min_4}\le h\le h^{\max_4}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} T=T^{\max_4}(t)\\ \mathit{as\ long\ as}\ h^{\min_4}\le h\le h^{\max_4};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_4}\ \mathit{or}\ h>h^{\max_4}\end{matrix}\right\}\\
\text{Min:}\\
\left\{\begin{matrix} T>T^{\min_4}(t)\\ \mathit{and}\\ h^{\min_4}\le h\le h^{\max_4}\end{matrix}\right\} \mathit{else} \left\{\begin{matrix} T=T^{\min_4}(t)\\ \mathit{as\ long\ as}\ h^{\min_4}\le h\le h^{\max_4};\\ Q_{n_T}=0\ \mathit{if}\ h<h^{\min_4}\ \mathit{or}\ h>h^{\max_4}\end{matrix}\right\}
\end{cases}
\end{array}
\tag{6.65}
$$

where

$$Q_{n_T} = -\int q_{n_T}\, d\Gamma \tag{6.66}$$

represents the integral boundary balance heat flux summed-up at discrete (nodal) points to which the corresponding boundary values are related (cf. Sect. 8.19.2), $(\ldots)^{\max_i}$ and $(\ldots)^{\min_i}$ denote the prescribed maximum and minimum bounds, respectively, for the corresponding type of BC and SPC, and T and h in (6.65) are the temperature and the hydraulic head, respectively, computed on the boundary or the singular point. The min-max bounds for the heat flux Q_{n_T}, the temperature T and the hydraulic head h can be again time-dependent functions.

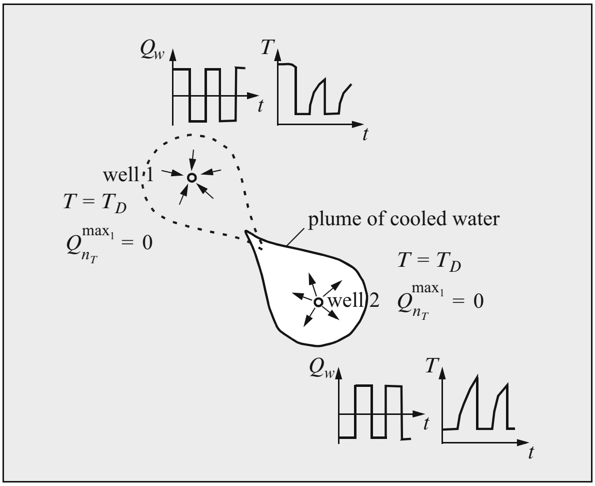


Fig. 6.10 Intermittent pumping regime of a well doublet system for heat extraction and re-injection (*horizontal view*)

Similar to the mass flux constraints in Sect. 6.4.2 the balance-based evaluation of heat fluxes must be exclusively related to the advective (convective) part

$$Q_{n_T} \approx -\int (T\, q_{n_h}) d\Gamma \tag{6.67}$$

to assure unambiguously directional balance quantities. An example of using BCC's and SPCC's for heat transport is schematized in Fig. 6.10 for a well doublet system under an intermittent pumping regime. The wells extract water from a heated aquifer in a time-given pumping operation $Q_w > 0$ for which the temperature at the wells has to be determined and re-inject cooled water with given temperature $T = T_D(t)$ as long as a recharging pumpage occurs $Q_w < 0$. Both wells comprise a temperature BC of 1st kind with $T = T_D(t)$ and a minimum heat flux constraint of zero $Q_{n_T}^{\min_1} = 0$.

6.5 Special BC's

6.5.1 Free (Phreatic) Surface BC

Free surface and phreatic surface are used as a synonym for porous-media problems describing the upper bound of a saturated zone (see Fig. 6.11 and definitions introduced in Sects. 2.2.1 and 2.2.2). A free (phreatic) surface is a moving boundary and subjected to two conditions: (1) a constant liquid pressure, usually taken to be $p = 0$ as the atmospheric pressure, and (2) a given mass conservation of flux across the macroscopic surface of discontinuity. The first pressure condition $p = \psi = 0$ is equivalent to $h = x_j$ expressed by the hydraulic head h, cf. (3.260), where x_j

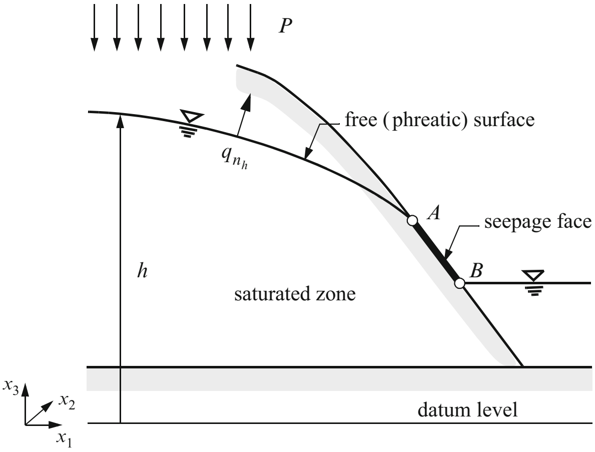


Fig. 6.11 Free (phreatic) surface and seepage face, $\widehat{AB}$

is the coordinate aligned to the gravity direction (e.g., vertical coordinate $x_3 = z$). The second condition is derived in Sect. 3.10.7 in form of (3.295). Both conditions finally lead to the following formulation of a free (phreatic) surface:

$$\left.\begin{aligned} q_{n_h} &= \varepsilon_e \frac{\partial h}{\partial t} - P \\ h &= x_j \end{aligned}\right\} \tag{6.68}$$

where ε_e is the specific yield (3.296) and P is the rate of infiltration (groundwater recharge). Note that for a pure (non-porous) liquid flow $\varepsilon_e = 1$. According to (6.68) the two BC's imposed on a free (phreatic) surface are to be satisfied simultaneously, viz.,

- A prescribed flux rate (as an infiltration or, if equal to zero, then impervious) as Neumann-type BC and
- The location corresponds to the hydraulic head, the water table (constant pressure level) as Dirichlet-type BC

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

6.5.2 Seepage Face BC

It is possible that a free surface approaches a rigid boundary of known geometry on which the flow can freely drain out the saturated porous-medium domain. Such a boundary is called a *seepage face* as illustrated in Fig. 6.11 for the boundary segment $\widehat{AB}$. The shape of the seepage face is known, except for the location of its end point A, which represents the point, where the *a priori* unknown free surface is terminated. Accordingly, the extent of a seepage face is initially unknown and its solution also leads to a nonlinear task.

Since a seepage face is exposed to the atmosphere, the condition $p = 0$ or equivalently $h = x_j$ must be imposed. Additionally, a seepage face only allows drainage, i.e., through it the liquid seeps *out*. This can be enforced by applying a constraint condition, where it is required that the balanced flux Q_{n_h} (6.56) on the boundary is only directed outward, i.e., $Q_{n_h} < 0$ (note that a negative Q_{n_h} means outflow). Thus, a seepage face is formulated by the following two conditions:

$$\left.\begin{aligned} h &= x_j \\ Q_{n_h} &< Q_{n_h}^{\max_1} = 0 \end{aligned}\right\} \tag{6.69}$$

Mathematically, a seepage face corresponds to a Dirichlet-type BC with $h = h_D = x_j$ which is combined with a maximum flux constraint $Q_{n_h}^{\max_1}$ equal to zero according to (6.55).

Alternatively, instead of a Dirichlet-type BC allowing a free drainage through the boundary, the pressure condition of the seepage face can be prescribed by a Cauchy-type BC, which provides a *limited* drainage. It reads

$$\left.\begin{aligned} q_{n_h} &= -\Phi(x_j - h) \\ Q_{n_h} &< Q_{n_h}^{\max_1} = 0 \end{aligned}\right\} \tag{6.70}$$

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

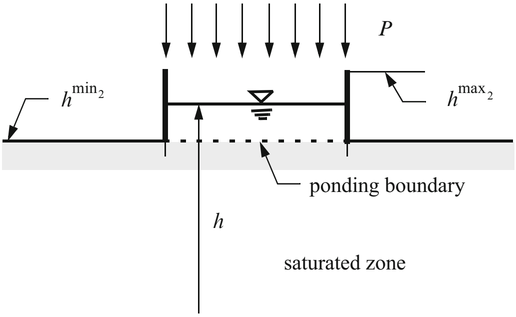
6.5.3 Surface Ponding BC

Surface ponding denotes a 'surface reservoir' BC to describe the storage of liquid (water) at the ground surface as illustrated in Fig. 6.12. This occurs when the liquid's pressure at ground surface satisfies the condition $p > 0$ (or $h > x_j = h^{\min_2}$). Usually, ponding is only allowed up to a maximum head, i.e., $h < h^{\max_2}$, where $h^{\max_2}(t)$ is a given maximum limit. Furthermore, mass conservation at the ponding boundary has to be imposed. Thus, the following formulation at a surface ponding boundary is required:

$$\left.\begin{aligned} q_{n_h} &= \frac{\partial h}{\partial t} - P \\ h^{\min_2} = x_j &< h < h^{\max_2} \end{aligned}\right\} \tag{6.71}$$

which is easily performed by a Neumann-type BC combined with min-max head constraints according to (6.55). Note that the first condition of (6.71) represents the interfacial mass conservation (3.295) for which the specific yield ε_e becomes unity (assuming that ponding on the ground surface occurs in an 'air layer'). Condition (6.71) can be recognized a specific free surface condition (6.68) which permits liquid to store on top of the ground.

Fig. 6.12 Surface ponding boundary



6.5.4 Integral BC

With respect to BC's of 2nd and 3rd kind special BC's are available for problems with free (phreatic) surface(s). They are referred to as *integral BC's* and are defined as follows:

$\int\times$ *2nd kind integral BC (integral Neumann type):*

$$\left.\begin{aligned}
&\text{Flow:}\\
&q_{n_h}(\boldsymbol{x},t) \;=\begin{cases} q_h(t) & \text{for 3D related to the initial stratigraphic structure}\\ \bar{q}_h(t) & \text{for 2D horizontal-unconfined as depth-integrated flux}\end{cases}\\
&\text{Mass:}\\
&q_{n_{kC}}(\boldsymbol{x},t) =\begin{cases} q_{kC}(t) & \text{for 3D related to the initial stratigraphic structure}\\ \bar{q}_{kC}(t) & \text{for 2D horizontal-unconfined as depth-integrated flux}\end{cases}\\
&\text{Heat:}\\
&q_{n_T}(\boldsymbol{x},t) \;=\begin{cases} q_T(t) & \text{for 3D related to the initial stratigraphic structure}\\ \bar{q}_T(t) & \text{for 2D horizontal-unconfined as depth-integrated flux}\end{cases}
\end{aligned}\right\} \tag{6.72}$$

$\int\otimes$ *3rd kind integral BC (integral Cauchy type):*

$$\left.\begin{aligned}
&\text{Flow:}\\
&q_{n_h}(\boldsymbol{x},t) \;=\begin{cases} -\Phi_h(h_C-h) & \text{for 3D related to the initial stratigraphic structure}\\ -\bar{\Phi}_h(h_C-h) & \text{for 2D horizontal-unconfined as depth-integrated flux}\end{cases}\\
&\text{Mass:}\\
&q_{n_{kC}}(\boldsymbol{x},t) =\begin{cases} -\Phi_{kC}(C_{kC}-C_k) & \text{for 3D related to the initial stratigraphic structure}\\ -\bar{\Phi}_{kC}(C_{kC}-C_k) & \text{for 2D horizontal-unconfined as depth-integrated flux}\end{cases}\\
&\text{Heat:}\\
&q_{n_T}(\boldsymbol{x},t) \;=\begin{cases} -\Phi_T(T_C-T) & \text{for 3D related to the initial stratigraphic structure}\\ -\bar{\Phi}_T(T_C-T) & \text{for 2D horizontal-unconfined as depth-integrated flux}\end{cases}
\end{aligned}\right\} \tag{6.73}$$

Using these integral formulations of flux BC's it is ensured that a given flux rate on their boundary portions becomes independent of the actually discharging aquifer thickness and the location of free surface. This is unlike a default nonintegral BC where a flux rate is integrated along the effective aquifer thickness, which depends on the actual (computed) free-surface position, and accordingly, varying (gross-) discharges may occur through such boundaries. As a result, it may happen that the total discharge through such varying boundaries constantly decreases at a descending water table. Consequently, such a flow region can inevitably fall dry and possibly the problem can 'collapse' with a zero inflow. Integral BC's prevent such situations since the gross discharges are not influenced by the location of free surface. The relation of fluxes, however, is distinguished in 2D and 3D applications due to reasons of implementation:

1. For 2D problems the fluxes have to be assigned as already depth-integrated. The dimension of these fluxes is then L^2T^{-1}, similar to a horizontal confined condition.
2. For 3D problems the aquifer system is compiled as an initial stratigraphic layer structure. BC's of the integral type are related to this initial structure and accordingly, the integrated gross discharges remain independent of the free-surface location during the computation with the BASD technique (see Sect. 9.5.3). Notice, the dimension of these boundary fluxes is LT^{-1} (not L^2T^{-1}).

Integral boundary flux conditions have only a distinct meaning for problems with free (movable) surface(s). If no free surfaces exist, they are totally equivalent to the nonintegral BC's of 2nd and 3rd kind, in accordance with (6.6), (6.19), (6.20), (6.37), (6.38) and (6.7), (6.22), (6.23), (6.40), (6.41), respectively.

6.5.5 Gradient-Type BC

Applied to unsaturated problems a Neumann flux-type BC (6.6) in the form

$$-\big[k_r(s)\boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})\big] \cdot \boldsymbol{n} = q_h \tag{6.74}$$

can be sometimes inappropriate, for instance if modeling a drainage boundary in the vadose zone with a bottom outflow BC for situations where the water table is located far below the domain of interest (Fig. 6.13). Here, a *gradient-type BC* is often to be preferred [362] written as

$$-\big\{\boldsymbol{K} f_\mu \cdot [\nabla \psi + (1+\chi)\boldsymbol{e}]\big\} \cdot \boldsymbol{n} = q_h^{\nabla} \tag{6.75}$$

On such a boundary it can be assumed that the pressure gradient diminishes $\nabla\psi \approx \mathbf{0}$ and (6.75) can be practically applied in the following form:

$$-\big\{\boldsymbol{K} f_\mu \cdot [(1+\chi)\boldsymbol{e}]\big\} \cdot \boldsymbol{n} = q_h^{\nabla} \tag{6.76}$$

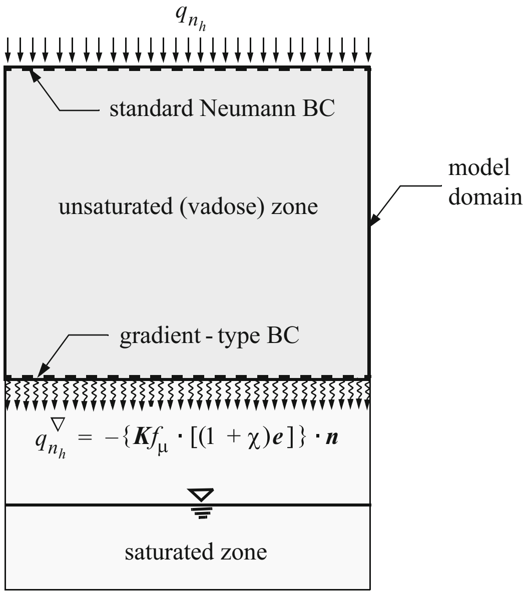

q_{n_h}
standard Neumann BC
model domain
unsaturated (vadose) zone
gradient - type BC
$q_{n_h}^{\nabla} = -\{\boldsymbol{K} f_{\mu} \cdot [(1+\chi)\boldsymbol{e}]\} \cdot \boldsymbol{n}$
saturated zone

Fig. 6.13 Gradient-type BC for free bottom outflow from a vadose zone with deep water table

Once $(1 + \chi)\boldsymbol{e} \cdot \boldsymbol{n} \neq 0$ the boundary freely drains the flow domain due to the influence of gravity.

6.5.6 Multilayer Well BC

The prescription of well-type BC in 3D heterogeneous aquifers under confined or unconfined conditions requires a more general formulation to model the effects of well bore storage and the vertical gradients of variables (hydraulic head, concentrations, temperature) along the well bore and well screens in a more realistic way. The standard well-type SPC's in form of (6.15), (6.32) and (6.50) are only applicable to singular points in the domain. Those points are per se not linked among each other and could not suitably present a well bore and well screen, where a relatively uniform distribution of a priori unknown head (or concentration and temperature) results from the high conductivity of the conduit that transmits flow, species mass and energy between different locations. Conventionally, iterative procedures (e.g., [384]) are used to adapt a uniform distribution of variables (e.g., hydraulic head h) at a series of points forming a well or well screen when mimicked via standard well-type SPC's. But, this technique is cumbersome and rather inefficient.

In contrast, the present *multilayer well BC* is a noniterative, straightforward, efficient and accurate method for handling well bore conditions in 3D aquifer systems which can consist of different layers or heterogeneous formations. Even in a 3D homogeneous aquifer, where a partially penetrating pumping well has to be

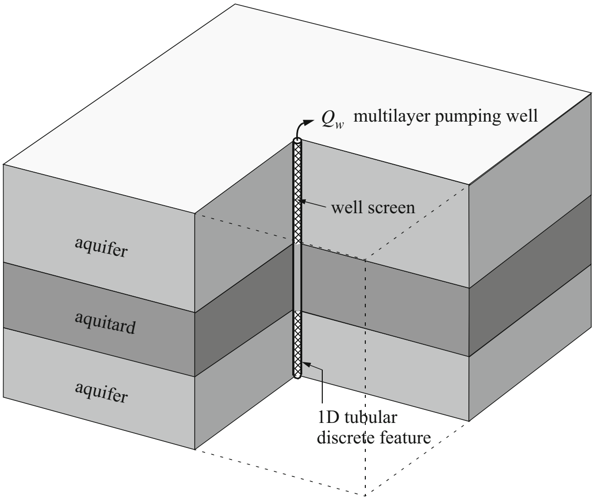


Fig. 6.14 Aquifer system containing a multilayer pumping well

imposed, the multilayer well BC is superior because the depth-variable inflow to the well is naturally accommodated.

The multilayer well BC involves a method, which superimposes high-conductivity 1D tubular discrete features (see Chap. 4) representing the well bore and well screens (Fig. 6.14). It was firstly introduced by Sudicky et al. [502] for aquifer flow problems and extended to contaminant transport by Lacombe et al. [328]. The use of high-conductivity 1D discrete features to represent a well ensures a uniform head (or concentration and temperature) along the well bore and well screens, with slight vertical gradients in the well toward the point where the well discharges. Storage in the well casing can also be accommodated by the superposition of the 1D discreate features. This effect can be significant at early times due to a rapid withdrawal of liquid from these features.

Assuming that the flow in the well along its axis is laminar and that the effect of storage in the well casing can be uniformly distributed along the length of the well bore, the 1D discrete feature equation describing transient liquid flow along the axis of the well bore is given according to Table 4.5, case TP, pure liquid:

$$\pi R^2\Big(\frac{1}{L_w} + \rho_0 g \gamma\Big)\frac{\partial h}{\partial t} - \pi R^2 K_w \frac{\partial}{\partial s}\Big[f_\mu\Big(\frac{\partial h}{\partial s} + \chi \boldsymbol{e}\Big)\Big] = -Q_w \delta(s - s_w) \qquad (6.77)$$

in which

$$K_w = \frac{R^2 \rho_0 g}{8\mu_0} \qquad (6.78)$$

by using the Hagen-Poiseuille law (4.51), where Q_w is the total pumping rate of the well, s is the arc length along the well bore (for vertical boreholes s is identical to vertical coordinate $x_3 = z$), s_w is the location of the point that is assigned to discharge (or recharge) the well bore, h is the hydraulic head in the well, L_w is the total length of the liquid-filled well bore, R is the radius of the well casing and screen(s), assuming to be equal, $\delta()$ is the Dirac delta function in 1D, γ is the compressibility of liquid, f_μ is the viscosity relation function of liquid (3.264), χ is the buoyancy coefficient (3.265), $\boldsymbol{e}$ is gravitational unit vector (3.261), g is the gravitational acceleration, ρ_0 is the reference density of liquid and μ_0 is the reference viscosity of liquid. Equation (6.77) is written for a well in which the casing is open (unconfined) to the atmosphere so that the storage in the well occurs due to a change in the water table and effects by compressibility of liquid.

Analogously, based on the derivations done in Chap. 4 and summarized in Tables 4.6 and 4.7, 1D discrete feature equations of the well bore can be formulated for species mass transport (Table 4.6, case TP, pure liquid)

$$\begin{aligned}\pi R^2 \frac{\partial C_k}{\partial t} + \pi R^2 v \frac{\partial C_k}{\partial s} - \pi R^2 \frac{\partial}{\partial s}\Big[\big(D_k + D_{(k)\mathrm{mech}}\big)\frac{\partial C_k}{\partial s}\Big] \\ + \pi R^2 \vartheta_k\, C_k = -(C_{kw} - C_k) Q_w \delta(s - s_w)\end{aligned} \qquad (6.79)$$

using Taylor's relation (4.69) of mechanical dispersion in a liquid-filled tube under laminar conditions as

$$D_{(k)\mathrm{mech}} = \frac{R^2 v^2}{48 D_k} \qquad (6.80)$$

where C_k is the concentration of species k in the well, v is the velocity of liquid in the well bore, D_k is the free-solution diffusion coefficient of species k, ϑ_k is the decay rate of species k and C_{kw} is the prescribed concentration of species k at well point s_w,

and for heat transport (Table 4.7, case TP, pure liquid)

$$\begin{aligned}\pi R^2 \rho c \frac{\partial T}{\partial t} + \pi R^2 \rho c v \frac{\partial T}{\partial s} - \pi R^2 \frac{\partial}{\partial s}\Big[\big(\Lambda + \rho c D_{\mathrm{mech}}\big)\frac{\partial T}{\partial s}\Big] \\ = -(T_w - T)\rho c Q_w \delta(s - s_w)\end{aligned} \qquad (6.81)$$

using solute-analogous Taylor's relation (4.69) for thermal mechanical dispersion in a liquid-filled tube under laminar conditions according to

$$D_{\mathrm{mech}} = \frac{R^2 v^2 \rho c}{48 \Lambda} \qquad (6.82)$$

where T is the temperature in the well, ρ is the density of liquid, c is the specific heat capacity of liquid, Λ is the coefficient of thermal conductivity of liquid and T_w is the prescribed temperature at well point s_w.

The governing equations (6.77), (6.79) and (6.81) for flow, species mass transport and heat transport, respectively, are formulated for a liquid-filled well bore tube. However, in cases, where the borehole is filled (or partially filled) with aquifer sediments (e.g., abandoned borehole), the well bore equations could be applied to porous-medium flow and transport conditions, which can be taken from Tables 4.5–4.7. Then, the K_w of (6.78) has to be replaced by Darcy's hydraulic conductivity and D_{mech} of (6.80) and (6.82) by the Scheidegger-Bear dispersion relation (4.68). More complex situations occur in heat transport for borehole heat exchanger (BHE), where different individual pipes and grout components are placed into a cylindrical borehole. The concept of multilayer BC must then be extended as further described in Sect. 13.5.

6.5.7 Outflow BC (OBC)

Often in mass and heat transport the liquid flows *through* (i.e., both into and out of) the computational domain Ω and advects transport quantities (concentration C_k, temperature T). This situation is necessitated by the fact that the true physical domain of interest is much too large to even be considered in a numerical simulation. Particularly, we have to consider outflow conditions in which the computational domain is *truncated* and suited BC's have to be necessarily applied at these 'artifical boundaries' of the truncated domain.

An outflow boundary of a truncated domain is often delicate to handle because the advective (convective) and dispersive quantities cannot be specified *a priori*. The goal of an outflow BC (OBC) is then to allow the transport quantities to leave freely with a minimal influence on the upstream solution. In practice, outflow boundaries are often subject to the assumption that the gradient of the transport quantity is zero (i.e., a common natural BC of Neumann type with $\nabla C_k = \mathbf{0}$ and/or $\nabla T = \mathbf{0}$), viz.,

$$\begin{aligned} -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= 0 \\ -(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} &= 0 \end{aligned} \tag{6.83}$$

with the consequence that the boundary is impermeable to the normal diffusive (dispersive/conductive) fluxes. The question arises how such a common natural BC does influence the solution upstream on the effluent boundary. To enlighten the situation let us consider a domain, which becomes truncated by a transition zone of infinitesimal thickness $\delta \to 0$ representing an outflow boundary Γ_+ as shown in Fig. 6.15 for heat transport. Providing the OBC in form of the transition zone has no conserved property, heat balance requires that the temperature T varies continuously and should not be changed by the presence of the boundary compared to the untruncated domain. Apparently, the boundary permeable to both the advective (convective) part $\rho c T \boldsymbol{q}$ and the conductive (dispersive) part $-\boldsymbol{\Lambda} \cdot \nabla T$ of the total heat flux $\boldsymbol{j}_T^\dagger$ permits upgradient heat movement by conduction. However, if the temperature gradient at the boundary is forced to zero, the conductive component of the heat flux is dropped at the boundary and the temperature profile differs over a certain distance upstream from the boundary as evidenced in Fig. 6.15. The

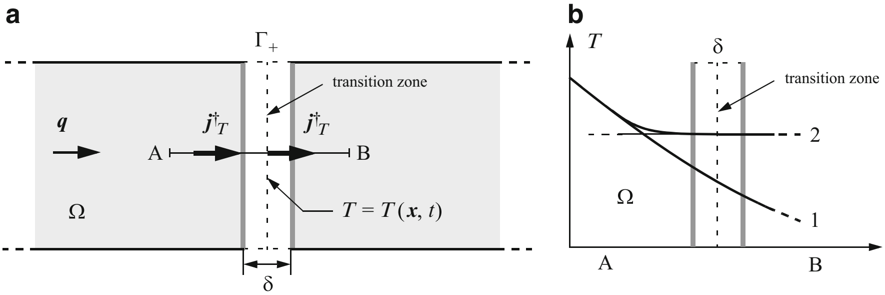


Fig. 6.15 Finite transition zone representation of outflow boundary Γ_+ of domain Ω: (**a**) continuity of total heat flux $\boldsymbol{j}_T^{\dagger} = \rho c T \boldsymbol{q} - \boldsymbol{\Lambda} \cdot \nabla T$ within the transition zone of infinitesimal thickness δ, where inside the temperature $T = T(\boldsymbol{x}, t)$ may vary continuously, (**b**) profile showing the behavior of temperature T when it varies continuously (1) and when temperature gradient is forced to zero (2) (Modified from [103])

measure of this upstreaming alteration in the temperature profile is controlled by the ratio between advection (convection) and conduction (dispersion). If advection dominates this alteration effect is usually small. On the other hand, if heat transport is dominated by thermal conduction, which possesses upstream conduction at the outflow boundary to a greater extent, a zero-gradient condition could not be a good choice. The situation can be mitigated if we can choose a more appropriate location of the outflow boundary far enough, where the gradients are small or negligible during the simulation. Moreover, there are applications, where the zero-gradient condition is useful. For instance, the outlet into a big reservoir, where the temperature is perfectly mixed out.

We have to ask what is a better OBC than the common natural BC of Neumann type in form of (6.83). Alternative formulations have been analyzed by Gresho and Sani [209] in a numerical context. A promising OBC treatment is proposed by Frind [175] and Cornaton et al. [103] termed as *free exit BC* and *implicit Neumann condition*, respectively. It consists in the following: Instead of explicitly prescribing the Neumann-type BC's for mass and heat transport written in the convective form

$$\begin{aligned} q_{n_{kC}} &= -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} \\ q_{nT} &= -(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} \end{aligned} \tag{6.84}$$

and in the divergence form

$$\begin{aligned} q_{n_{kC}} &= C_k q_{nh} - (\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} \\ q_{nT} &= \rho c (T - T_0) q_{nh} - (\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} \end{aligned} \tag{6.85}$$

the boundary terms of (6.84) and (6.85) are treated as unknown quantities and put back onto the LHS for the numerical solution. In this way, no assumptions must be made anymore for the gradients of the concentration or temperature. This form of OBC ensures that mass and heat fluxes become freely permeable at the boundary both to the advective (convective) and dispersive (conductive) components of transport. For the divergence forms of transport the OBC needs the knowledge of the advective flux $q_{nh} = \boldsymbol{q} \cdot \boldsymbol{n}$ at the outflow boundary. In general, q_{nh} is a prior unknown and must be determined from the flow equation via a postprocessing balance analysis. The numerical treatment of OBC's is described in Sects. 8.5.3 and 8.9.

Chapter 7
Anisotropy

7.1 Principal Directions and Rotation

For the tensors of hydraulic conductivity $\boldsymbol{K}$ (3.263), transmissivity $\boldsymbol{T}$ (3.302) and thermal conductivity of solid $\boldsymbol{\Lambda}_0^s$ (3.172) anisotropy is to be taken into account. It is assumed they have *orthotropic* properties, i.e., the conductivities are given along their principal directions x_i^m $(i = 1, 2, 3)$, which are turned against the global Cartesian coordinate system x_i $(i = 1, 2, 3)$ by a rigid-body rotation (Fig. 7.1), see Sect. 2.1.5.3. Hence, the conductivity along the principal directions x_i^m represents a diagonal matrix

$$K_{ij}^m = \begin{pmatrix} K_1^m & 0 & 0 \\ 0 & K_2^m & 0 \\ 0 & 0 & K_3^m \end{pmatrix} \tag{7.1}$$

which can be mapped by using the transformation in form of the rotation matrix A_{ij} (2.57)

$$A_{ij} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \tag{7.2}$$

onto the global Cartesian coordinate system x_i, viz.,

$$K_{ij} = A_{li}\, K_{lk}^m\, A_{kj} \tag{7.3}$$

or

$$K_{ij} = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{pmatrix} \begin{pmatrix} K_1^m & 0 & 0 \\ 0 & K_2^m & 0 \\ 0 & 0 & K_3^m \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \tag{7.4}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_7,

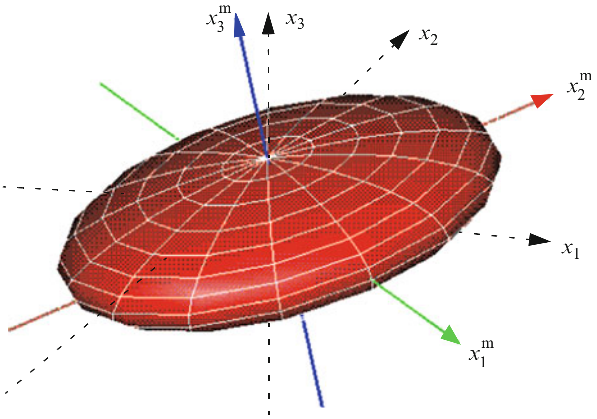


Fig. 7.1 Definition of a 3D anisotropic conductivity

After rotation, the tensor $\boldsymbol{K}$ in the global Cartesian coordinates $\boldsymbol{x}$ results, which describes an anisotropic conductivity in the form:

$$K_{ij} = \begin{pmatrix} A_{11}^2 K_1^m + A_{21}^2 K_2^m + A_{31}^2 K_3^m & A_{11}A_{12}K_1^m + A_{21}A_{22}K_2^m + A_{31}A_{32}K_3^m & A_{11}A_{13}K_1^m + A_{21}A_{23}K_2^m + A_{31}A_{33}K_3^m \\ & A_{12}^2 K_1^m + A_{22}^2 K_2^m + A_{32}^2 K_3^m & A_{12}A_{13}K_1^m + A_{22}A_{23}K_2^m + A_{32}A_{33}K_3^m \\ \text{symm.} & & A_{13}^2 K_1^m + A_{23}^2 K_2^m + A_{33}^2 K_3^m \end{pmatrix} \tag{7.5}$$

The elements of the rotation matrix (7.2), A_{ij}, represent the directional cosines of the rotation angles (2.60) appearing in transforming from the Cartesian coordinate system x_i to the principal direction system x_i^m. They are referred to as *Eulerian angles* [194] as defined in Fig. 2.7. In 3D three angles (ϕ, θ, ψ) result, while in 2D only one Eulerian angle ϕ has to be specified to perform a unique principal axis rotation. As the result, anisotropy for 3D applications needs formally six parameters $(K_1^m, K_2^m, K_3^m, \phi, \theta, \psi)$, while three parameters (K_1^m, K_2^m, ϕ) are required to handle a 2D anisotropy. Considering the effort and the appropriateness in computing the anisotropic conductivity (7.5), different strategies are employed for 3D and 2D cases as described subsequently.

7.2 Two-Dimensional Anisotropy

Two-dimensional anisotropy represents a special case of the principal direction transformation as introduced above. It rotates the principal axes of the conductivity tensor K_{ij}^m into the axes of the global Cartesian coordinate system (x_1, x_2) by a

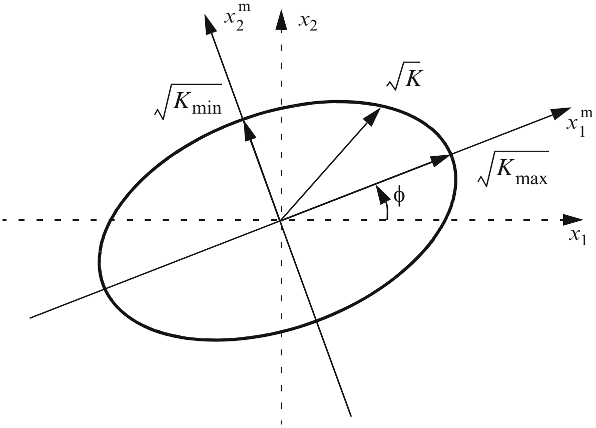


Fig. 7.2 Definition of a 2D anisotropic conductivity

given angle ϕ between the first principal axis of K_{ij}^m and the x_1–axis (Fig. 7.2). An anisotropic conductivity is then uniquely defined by their maximum and minimum conductivities, $K_{\max} \equiv K_1^m$ and $K_{\min} \equiv K_2^m$, respectively, acting along the principal axes, and by the rotation angle ϕ, as depicted in Fig. 7.2.

The components of the conductivity tensor K_{ij} in the global coordinate system (x_1, x_2) are determined by using the rotation formula (7.4) for the 2D case. Expressing the directional cosines A_{ij}

$$A_{ij} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \tag{7.6}$$

by the rotation angle ϕ it yields:

$$K_{ij} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} K_{\max} & 0 \\ 0 & K_{\min} \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \tag{7.7}$$

The components of the 2D anisotropic conductivity K_{ij} lead finally to:

$$\begin{aligned} K_{11} &= K_{\max}\cos^2\phi + K_{\min}\sin^2\phi \\ K_{22} &= K_{\max}\sin^2\phi + K_{\min}\cos^2\phi \\ K_{12} &= K_{21} = (K_{\max} - K_{\min})\,\sin\phi\,\cos\phi \end{aligned} \tag{7.8}$$

In FEFLOW for each discretized (finite) element following quantities are input:

1. The maximum conductivity $K_{\max}$.
2. A ratio of anisotropy defined as:

$$\Xi_{\text{aniso}} = K_{\min}/K_{\max} \quad \text{with} \quad K_{\max} > 0 \tag{7.9}$$

being unity for isotropic relations.
3. The rotation angle ϕ.

Analogous relationships exist for the tensor of transmissivity $\boldsymbol{T}$ (3.302) applied to 2D problems of confined aquifers. On the other hand, thermal conductivities of solid in 2D are assumed isotropic $\boldsymbol{\Lambda}_0^s = \Lambda^s \boldsymbol{\delta}$, (3.172), see Sect. 7.4.2.

7.3 Three-Dimensional Anisotropy

7.3.1 General 3D Anisotropy Formulation

The 3D rotation of the principal axes needs the knowledge of the three Eulerian angles (ϕ, θ, ψ), which are defined in Fig. 2.7. If the Eulerian angles are explicitly given, the directional cosines A_{ij} of the rotation matrix (7.2) can be immediately expressed as [194]:

$$A_{ij} = \begin{pmatrix} \cos\psi\cos\phi - \cos\theta\sin\phi\sin\psi & \cos\psi\sin\phi + \cos\theta\cos\phi\sin\psi & \sin\psi\sin\theta \\ -\sin\psi\cos\phi - \cos\theta\sin\phi\cos\psi & -\sin\psi\sin\phi + \cos\theta\cos\phi\cos\psi & \cos\psi\sin\theta \\ \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{pmatrix} \tag{7.10}$$

This represents the most general formulation of anisotropy in 3D. This case is optionally available in FEFLOW, where the user can directly input the six parameters $(K_1^m, K_2^m, K_3^m, \phi, \theta, \psi)$.

The general rotation matrix (7.10) contains the following three important special cases of axis rotation:

(i) *Rotation about the x_3−axis only:*

$$A_{ij} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{at} \quad \theta = \psi = 0^{\text{o}} \tag{7.11}$$

(ii) *Rotation about the x_2−axis only:*

$$A_{ij} = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \quad \text{at} \quad \begin{matrix} \phi & = -90^{\text{o}} \\ \psi & = \ \ 90^{\text{o}} \end{matrix} \tag{7.12}$$

(iii) *Rotation about the x_1−axis only:*

$$A_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \quad \text{at} \quad \phi = \psi = 0^{\text{o}} \tag{7.13}$$

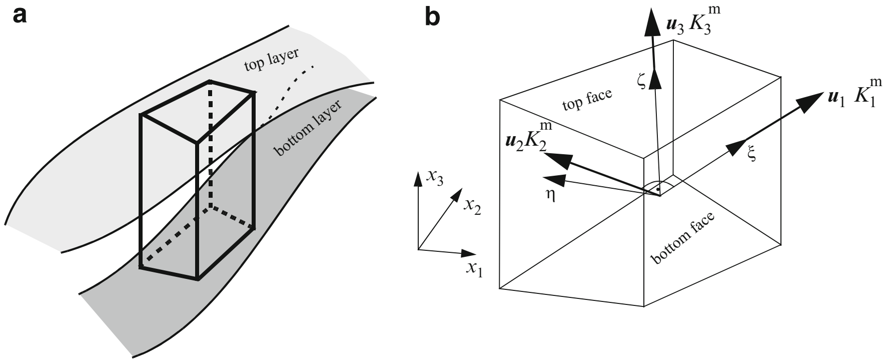


Fig. 7.3 Layer-oriented principal directions of anisotropy for a prismatic finite element

7.3.2 Shape-Derived 3D Anisotropy by Springer's Method

Applying the rotation matrix (7.10) to the transformation (7.5) it leads to relatively expensive formulae for the resulting conductivity matrix K_{ij} in the 3D anisotropic case. Unlike the 2D case, where the additional effort in prescribing only one Eulerian angle ϕ for the definition of an anisotropic property of conductivity is still justifiable, in 3D the Eulerian angles (ϕ, θ, ψ) are often not utilized as a direct input because, in practice, they are generally not explicitly known and not given.

On the other hand, one can make use of the fact that the principal axes generally correlate with the geologic layer structure and, accordingly, provided that the 3D shape of the layers is known, the spatial rotation of the principal directions is automatically accomplished via computation. The conductivity in a geologically layered structure is often orthotropic inasmuch the higher conductivity $K_1^m = K_2^m$ is parallel to the layering and a lower value K_3^m is normal to the stratigraphy. It leads to a computational approach in the FEM (cf. Sect. 8.11) for transforming the principal axes of each finite element if 3D prismatic finite elements are used, which are capable of fitting their top and bottom faces to the actual stratigraphic layering (Fig. 7.3a).

For 3D prismatic finite elements a method of shape-derived principal directions has been proposed by Springer [489]. Using quadrilateral or triangular prismatic elements the two opposite top and bottom faces of each element represent layer boundaries. A finite element is usually characterized by its local distorted coordinate system (ξ, η, ζ) as shown in Fig. 7.3b. The coordinate transformation from the local system (ξ, η, ζ) to the global Cartesian system (x_1, x_2, x_3) is achieved by the Jacobian matrix $\boldsymbol{J}$ (2.44):

$$\boldsymbol{J} = \begin{pmatrix} x_{1,\xi} & x_{2,\xi} & x_{3,\xi} \\ x_{1,\eta} & x_{2,\eta} & x_{3,\eta} \\ x_{1,\zeta} & x_{2,\zeta} & x_{3,\zeta} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial x_1}{\partial \xi} & \dfrac{\partial x_2}{\partial \xi} & \dfrac{\partial x_3}{\partial \xi} \\ \dfrac{\partial x_1}{\partial \eta} & \dfrac{\partial x_2}{\partial \eta} & \dfrac{\partial x_3}{\partial \eta} \\ \dfrac{\partial x_1}{\partial \zeta} & \dfrac{\partial x_2}{\partial \zeta} & \dfrac{\partial x_3}{\partial \zeta} \end{pmatrix} \tag{7.14}$$

The idea behind Springer's method of shape-derived principal directions is in relating appropriately the three principal axes of the conductivity anisotropy to the local finite-element coordinate system (ξ, η, ζ). Basically, the directions of the mutually orthogonal principal axes are identified by the following normalized three vectors $\boldsymbol{u}_i$ $(i = 1, 2, 3)$ (Fig. 7.3b):

$$\boldsymbol{u}_1 = \begin{pmatrix} x_{1,\xi} \\ x_{2,\xi} \\ x_{3,\xi} \end{pmatrix} \quad \text{(parallel to } \xi\text{−direction)} \tag{7.15}$$

$$\begin{gathered} \boldsymbol{u}_2 = m_{11} \begin{pmatrix} x_{1,\eta} \\ x_{2,\eta} \\ x_{3,\eta} \end{pmatrix} - m_{12} \begin{pmatrix} x_{1,\xi} \\ x_{2,\xi} \\ x_{3,\xi} \end{pmatrix} \\ \text{(orthogonalization of } \begin{pmatrix} x_{1,\eta} \\ x_{2,\eta} \\ x_{3,\eta} \end{pmatrix} \text{ with respect to } \begin{pmatrix} x_{1,\xi} \\ x_{2,\xi} \\ x_{3,\xi} \end{pmatrix} \text{)} \end{gathered} \tag{7.16}$$

$$\begin{gathered} \boldsymbol{u}_3 = \begin{pmatrix} x_{1,\xi} \\ x_{2,\xi} \\ x_{3,\xi} \end{pmatrix} \times \begin{pmatrix} x_{1,\eta} \\ x_{2,\eta} \\ x_{3,\eta} \end{pmatrix} \\ \text{(perpendicular to the area spanned by } \begin{pmatrix} x_{1,\xi} \\ x_{2,\xi} \\ x_{3,\xi} \end{pmatrix} \text{ and } \begin{pmatrix} x_{1,\eta} \\ x_{2,\eta} \\ x_{3,\eta} \end{pmatrix} \text{)} \end{gathered} \tag{7.17}$$

where

$$m_{ij} = \boldsymbol{J} \cdot \boldsymbol{J}^T = \begin{pmatrix} x_{1,\xi}^2 + x_{2,\xi}^2 + x_{3,\xi}^2 & x_{1,\xi}x_{1,\eta} + x_{2,\xi}x_{2,\eta} + x_{3,\xi}x_{3,\eta} & x_{1,\xi}x_{1,\zeta} + x_{2,\xi}x_{2,\zeta} + x_{3,\xi}x_{3,\zeta} \\ & x_{1,\eta}^2 + x_{2,\eta}^2 + x_{3,\eta}^2 & x_{1,\eta}x_{1,\zeta} + x_{2,\eta}x_{2,\zeta} + x_{3,\eta}x_{3,\zeta} \\ \text{symm.} & & x_{1,\zeta}^2 + x_{2,\zeta}^2 + x_{3,\zeta}^2 \end{pmatrix} \tag{7.18}$$

As the result, the directional cosines can be simply expressed by

$$A_{ij} = \cos(\boldsymbol{u}_i, \boldsymbol{e}_j) = \frac{\boldsymbol{u}_i \cdot \boldsymbol{e}_j}{\|\boldsymbol{u}_i\| \|\boldsymbol{e}_j\|} \quad (i, j = 1, 2, 3) \tag{7.19}$$

with the base vectors (2.5) in 3D:

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{7.20}$$

to compute the rotation matrix A_{ij}. Finally, it yields:

$$A_{ij} = \begin{pmatrix} \frac{x_{1,\xi}}{\sqrt{m_{11}}} & \frac{x_{2,\xi}}{\sqrt{m_{11}}} & \frac{x_{3,\xi}}{\sqrt{m_{11}}} \\ \frac{m_{11}x_{1,\eta}-m_{12}x_{1,\xi}}{\sqrt{m_{11}^2 m_{22}-m_{11}m_{12}^2}} & \frac{m_{11}x_{2,\eta}-m_{12}x_{2,\xi}}{\sqrt{m_{11}^2 m_{22}-m_{11}m_{12}^2}} & \frac{m_{11}x_{3,\eta}-m_{12}x_{3,\xi}}{\sqrt{m_{11}^2 m_{22}-m_{11}m_{12}^2}} \\ \frac{x_{2,\xi}x_{3,\eta}-x_{2,\eta}x_{3,\xi}}{\sqrt{m_{11}m_{22}-m_{12}^2}} & \frac{x_{3,\xi}x_{1,\eta}-x_{3,\eta}x_{1,\xi}}{\sqrt{m_{11}m_{22}-m_{12}^2}} & \frac{x_{1,\xi}x_{2,\eta}-x_{1,\eta}x_{2,\xi}}{\sqrt{m_{11}m_{22}-m_{12}^2}} \end{pmatrix} \tag{7.21}$$

In the practical simulation, the rotation matrix is computed by using (7.21) at each Gaussian integration point of a finite element according to the 3D stratigraphic layer structure. In this way, the anisotropic 3D tensor of conductivity K_{ij} is automatically determined by applying (7.5).

The key advantages of this method concern that only the conductivities of the three principal directions along the layer structure need to be input for each discretized element, viz.,

1. The conductivity K_1^m parallel to layering,
2. The conductivity K_2^m parallel to layering, and
3. The conductivity K_3^m normal to layering,

and the 3D shape of the layer geometry is embodied in the anisotropy relations with a relatively high accuracy.

7.4 Special Cases

In specifying the hydraulic conductivity matrix $\boldsymbol{K}$ and the thermal conductivity $\boldsymbol{\Lambda}_0^s$ of solid two special cases are important:

7.4.1 Axis-Parallel Anisotropy

The system of principal axes x_i^m is parallel to the Cartesian coordinate system x_i. Hence, the Eulerian angles become zero: $\phi = \theta = \psi = 0^{\circ}$ and the hydraulic conductivity matrix K_{ij} reduces to the diagonal matrix K_{ij} (7.1), viz.,

$$\boldsymbol{K} = K_{ij} = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix} = \begin{pmatrix} K_1^m & 0 & 0 \\ 0 & K_2^m & 0 \\ 0 & 0 & K_3^m \end{pmatrix} \tag{7.22}$$

or

$$K_{ii} = K_i^m \quad \text{and} \quad K_{ij} = 0 \quad \text{for} \quad i \neq j \quad \text{at} \quad i = 1,2,3 \tag{7.23}$$

This case is to be referred to as *axis-parallel anisotropy*. Here, different conductivities can be prescribed in each direction of the x_i−axes:

$$\begin{array}{ll} \text{For 3D:} & K_{11} = K_1^m, \quad K_{22} = K_2^m, \quad K_{33} = K_3^m \\ \text{For 2D:} & K_{\max} = K_{11} \quad \text{and} \quad \Xi_{\text{aniso}} = K_{22}/K_{11} \quad \text{with} \quad \phi = 0^{\text{o}} \end{array} \tag{7.24}$$

For the thermal conductivity $\boldsymbol{\Lambda}_0^s$ of solid a special form of axis-parallel anisotropy is applied to 3D problems

$$\boldsymbol{\Lambda}_0^s = \begin{pmatrix} \Lambda^s & 0 & 0 \\ 0 & \Lambda^s & 0 \\ 0 & 0 & \Lambda_3^s \end{pmatrix} \tag{7.25}$$

where only the thermal conductivity Λ_3^s of solid in the x_3−direction differs. Introducing the *thermal anisotropy factor* as

$$\Xi_{\text{aniso}}^{\Lambda} = \frac{\Lambda_3^s}{\Lambda^s} \tag{7.26}$$

the 3D thermal conductivity $\boldsymbol{\Lambda}_0^s$ of solid becomes

$$\boldsymbol{\Lambda}_0^s = \Lambda^s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \Xi_{\text{aniso}}^{\Lambda} \end{pmatrix} \tag{7.27}$$

7.4.2 Isotropy

The directional independency of the conductivity leads to an *isotropic conductivity* matrix K_{ij} in the form

$$K_{ij} = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix} = \begin{pmatrix} K & 0 & 0 \\ 0 & K & 0 \\ 0 & 0 & K \end{pmatrix} \tag{7.28}$$

or

$$\boldsymbol{K} = K\,\boldsymbol{\delta} \tag{7.29}$$

where K corresponds to an isotropic hydraulic conductivity coefficient. Here, the following is to be input:

$$\begin{array}{ll} \text{For 3D:} & K_{11} = K_{22} = K_{33} = K \\ \text{For 2D:} & K_{\max} = K \quad \varXi_{\text{aniso}} = 1 \quad \text{with} \quad \phi = 0^{\text{o}} \end{array} \tag{7.30}$$

Similarly, for the thermal conductivity $\boldsymbol{\varLambda}_0^s$ of solid the isotropic relation holds

$$\boldsymbol{\varLambda}_0^s = \varLambda^s\,\boldsymbol{\delta} \tag{7.31}$$

and only a scalar thermal conductivity $\varLambda^s$ of solid is input both in 2D and 3D problems with $\varXi_{\text{aniso}}^{\varLambda} \equiv 1$.

Part II
Finite Element Method

Part II is organized into eight chapters, which covers the finite element solution of the underlying flow, mass and heat transport equations in porous and fractured media derived in Part I. We start with a comprehensive introduction into the finite element method (FEM) applied to multiple dimensions. Its relationship to other numerical methods, such as finite difference method, finite volume method, spectral method and others, is discussed. At first, in Chap. 8 the basic principles of FEM are systematically developed and reviewed for prototypical advection-dispersion equations (ADE's). This chapter is clearly most important because the following chapters, in which the specific finite element solutions are elaborated for selected classes of problems, will adjunct to this one. The different spatial and temporal discretization techniques are addressed. The important approximate solutions for the divergence and convective forms of ADE are carefully developed. Emphasis is given on adaptive solution strategies. Implicit and explicit time integration methods are reviewed and compared. It clearly shows the superiority of implicit strategies for the present problem classes, in particular automatic error-controlled predictor-corrector schemes are favorited. Upwind methods are thoroughly discussed and examined in comparison to the standard Galerkin-based FEM (GFEM). The optimality of GFEM is explicitly shown in Appendix F. Stability and error analyses for the favorite schemes are presented in some detail. Their most practical outcome is summarized in Table 8.9 containing quantitative estimates of spurious numerical dispersion and stability bounds in space and time. Techniques for solving the resulting matrix equations are discussed. They cover direct Gaussian-based methods and various preconditioned iterative methods, such as conjugate gradient, ORTHOMIN, GMRES, bi-conjugate gradient stabilized and multigrid techniques. Of important interest is the solution of the nonlinear equations by using Picard and Newton iteration techniques, which are embedded in adaptive time stepping strategies for solving transient problems. A particular focus is given on derived quantities, i.e., the evaluation of fluxes and balance quantities. It is shown that the FEM is locally conservative.

The following chapters deal with the finite element solutions for saturated porous media (groundwater), variably saturated porous media (unsaturated-saturated flow), variable-density flow and transport in porous media, multispecies and single-species mass transport in reacting and non-reacting porous media, heat transport processes in porous media and discrete feature modeling for flow, mass and heat transport. Each class of problems also comprises typical examples and benchmark tests to illustrate the usefulness, efficiency and accuracy of the proposed numerical techniques in comparison to analytical solutions (if exist), physical measurements (if available) and other numerical findings (if precious). It provides a comprehensive overview of appropriate approaches for the different problems having their specific aspects and numerical requirements which should be very valuable for users who want to verify solutions or expand their modeling skills. To allow repeatability and individual rerun (either by using FEFLOW or with other programs), each example is fully documented and contains a complete description of the dataset summarized in tables. The examples are prototypical for more complex solutions. In the present book, it is not intended to present field applications which would restrict the comparability and the emphasis on essential features we like to highlight.

There are a number of topics which are reviewed in the individual chapters. Of particular concern are the treatment of free-surface problems, fully anisotropic flow situations, the incorporation of multi-layer well condition, the different formulations of Richards' equations with the favorite solution strategies suited for variably saturated flow, including the computation of hysterestic effects and time-varying porosity, the simulation of variable-density flow with analysis of important convection phenomena, including free convection and double-diffusive (thermohaline) convection, required schemes to tackle successfully buoyant flow and multispecies reactive mass transport, heat transport modeling including borehole heat exchanger, discrete feature modeling for flow, mass and heat transport with and without buoyancy effects as well as accurate budget analyses for flow, mass and heat. In a final chapter meshing strategies for finite elements, particle tracking techniques, useful methods of streamline computation and finite element interpolation are addressed.

Chapter 8
Fundamental Concepts of Finite Element Method (FEM)

8.1 Introduction

In the previous Chaps. 3–5 the governing continuum balance equations in form of partial differential equations (PDE's) have been derived for a wide range of flow, mass and heat transport processes in porous and fractured media. Their solution under given IC's and BC's, such as described in Chap. 6, requires appropriate and efficient mathematical methods, which can be firstly grouped into *analytical* and *numerical* methods. There is a family of powerful analytical methods (e.g., Fourier and Laplace transformation, complex variable techniques, Green's functions, perturbation methods, power series), which are capable of solving a certain number of problems in an exact way. However, exact analytical solutions are often only attainable for elementary linear (or quasi-linear) problems on simple (regular) geometries. Very few analytical solutions exist for nonlinear problems with regions of regular geometry, however, these are usually approximate solutions in terms of an infinite series or some transcendental functions that can be evaluated only approximately. If exact analytical solutions are available on idealized problems they are often advantageous in comparison to numerical results for purposes of verification and estimation of errors arising in the alternative numerical methods.

Problems involving irregular geometry, materials with variation in properties, nonlinear relationships and/or complex BC's are intractable by analytical methods and numerical methods must be used in general. They allow the solution for a broad range of problems. The key feature of any numerical method is in the *approximate* solution of the basic PDE's via spatial and temporal discretizations, in which the solution variables, which are basically continuous functions of space and time, are obtained by *discrete* values, defined at specific points in space and time (Fig. 8.1). In doing this approximation, the governing PDE's are replaced by a number (often, a very large number) of linear (or linearized) *algebraic* equations, which can be easily solved via computers. As a consequence of the numerical approximation, *errors* are naturally inherent in the solution and the big challenge of numerical methods is to minimize these numerical errors and find best accurate, convergent and stable

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_8,

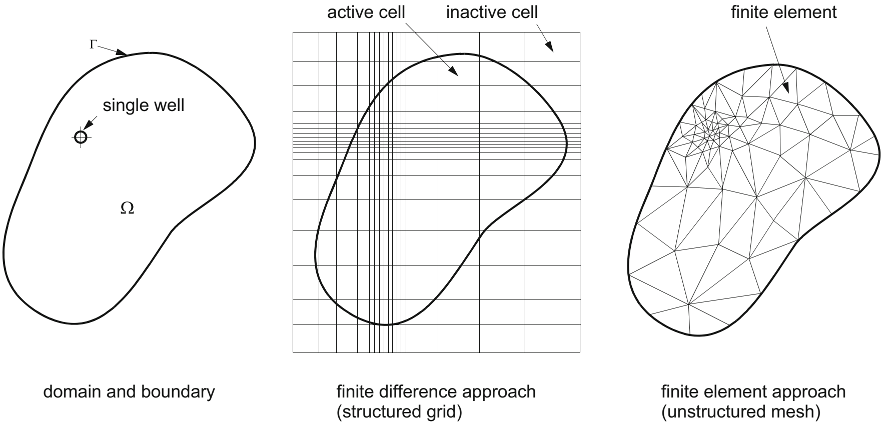


Fig. 8.1 Example of 2D domain discretized by finite differences and finite elements

solutions by using efficient, general and robust strategies of approximation. It is important to ensure that the approximation satisfies certain important properties of the exact solution, e.g., conservativity, boundedness and consistency (see Sect. 1.2.2 for further discussion).

We can classify the numerical methods as follows:

- Finite difference method (FDM)
- Method of characteristics (MOC)
- Finite element method (FEM)
- Finite volume method (FVM)
- Boundary element method (BEM)
- Meshless method (MLM)
- Spectral element method (SEM)

These methods are closely related. The FDM is the classic numerical approach, e.g., [168]. It is conceptually straightforward and had a high popularity in past. FDM approximates the differential form of the basic PDE in a difference form and is usually restricted to simple (rectangular) geometries and BC's. The specific advantage of FDM lies in the use of regular grids on which the approximation can be most efficiently performed. The development of finite-difference approximations is commonly done by either Taylor series expansion or curve-fitting technique.

The MOC as a traditional solution method [165,184] is only applicable to PDE of hyperbolic type, i.e., for advection-dominated transport processes. It is based on the concept of trajectories (or characteristics) on which a large number mathematical particles are tracked. While mainly 1D and partly 2D unsteady flow processes could be successfully modeled, the method is rather cumbersome when extended to multidimensional problems, dealing with complex BC's and nonlinearities.

The basic ideas underlying the FEM have a long history. Ritz [445] and Galerkin [181] presented variational integral formulations of a PDE and approximate solutions based on their minimization. Pioneering work of FEM in the modern form that we know today dates back to the early 1940s given by Hrenikoff [264] and Courant [104]. First applications were done for aero structures in the late 1950s [525]. Clough [88] coined the term *finite element method* at that time. The power of FEM was quickly recognized and the first textbook on FEM appeared in the mid-1960s by Zienkiewicz and Cheung [589], which boosted the development of FEM in many fields of sciences and engineering lasting up to now. First applications of FEM for porous-media problems were given by Zienkiewicz and Cheung [589], Pinder and Gray [421] and Huyakorn and Pinder [280]. Since then, the FEM has become one of the basic tools for numerical analysis in structural mechanics, fluid dynamics, heat transfer and numerical mathematics (for literature review see Sect. 1.3 with Table 1.3).

Today, the FEM represents a collection of theory-rich techniques and is based on the weak (or variational) formulation of the governing initial-boundary-value problem. This theoretical foundation on weak formulation is quite distinct from FDM. The weak formulation is an integral approach, which is a natural and an adequate approach of a continuum balance statement. FEM subdivides the continuum in a finite number of elements, for which the balance statements are discretely applied. The resultant algorithm of the FEM can be universally expressed as a matrix statement with all formation processes on a *generic* master element. The generic master element statement is then *assembled* into a global matrix statement. BC's can be brought directly into the generic master element providing accurate expressions of surface integrals for the PDE global domain boundary on which any flux-type BC is applicable. The FEM is essentially geometry-free. In principle, FEM can be applied to domains of arbitrary shape and with quite arbitrary BC's. FEM by its nature leads to unstructured meshes (Fig. 8.1). Most complex types of geometries can be simply handled. These features make the FEM a general, systematic, very powerful and highly flexible numerical method, which is superior to the other numerical methods.

There is a wide variety of methods called finite volume methods (FVM's), e.g., [83, 162]. Sometimes they are termed as control volume methods or previously, integrated FDM. FVM is usually also based on weak formulations of the basic problem similar to FEM, however, the approximation of the balance terms relies on evaluation of surface integrals, where boundary fluxes are developed via finite differences. In this process, the conservation is enforced across the surfaces of the adjoining control volume. It allows the construction of cost-effective schemes for both structured and unstructured grids. It has been demonstrated [209] that the FVM is inherently a FEM if using low-order elements (basically linear). It can be shown [83, 165] that FVM can be formulated from either FDM or FEM. Identical discrete schemes result for FVM and FEM [284] if using low-order approximations and equivalent meshing via control volumes and elements, respectively. However, serious problems with FVM can arise when cross-derivatives (such as anisotropic problems, e.g., associated with the hydrodynamic dispersion tensor $\boldsymbol{D}_k$, (3.184)) appear in the governing PDE. Commonly, diffusive/conductive gradient terms are approximated in FVM by using a two-point flux approximation (TPFA) scheme applied to two adjacent cell values. But, TPFA is insufficient to express diffusive fluxes, where off-diagonal values in an anisotropic diffusion/conduction tensor exist (cf. Chap. 7). To circumvent this drawback multi-point flux approximation (MPFA) can be used [412], which, however, requires a nonlinear evaluation making FVM rather cumbersome and potentially less accurate. Moreover, higher-order approximations and complex geometries on arbitrarily unstructured meshing can lead to further difficulties in FVM.

The BEM is based on boundary integral equations in which only the boundaries of a domain are used to obtain approximate solutions, e.g., [54, 350]. It reduces the solution of the problem to one dimension less than the original problem (e.g., a 3D problem is solved by a 2D approximation), however, the resulting matrix systems are full, whereas the other numerical methods generally result in sparse matrices. The most serious aspect with BEM is that a fundamental solution (free space Green's function) of the PDE must be available, which commonly requires linear equations with constant coefficients (i.e., homogeneous materials). Thus, the application of BEM is limited to special problems.

For all numerical methods mentioned so far mesh configurations are required consisting of elements, cells or control volumes formed by connecting nodal points in a predefined manner. Unlikely, various methods have been developed which depend on finite number of points rather than meshes. They are called meshless methods (MLM's), finite point methods (FPM's) or element free Galerkin (EFG) methods, e.g., [44, 352]. Although most of the meshless methods have high computational cost as compared to FEM, they provide advantages for a certain class of problems such as moving boundaries, phase transformation, crack propagation and large deformation in solids as well as modeling of multiscale phenomena. The major advantage of MLM is the elimination of the need for mesh generation, which can be itself a difficult task. However, MLM's are not (yet) sophisticated enough for application in a general context. They often require background cells to improve numerical stability and accuracy so that in current practice, MLM's have shown not to be truly mesh-free.

The idea of MLM's has been adopted and modified in the so-called eXtended FEM (XFEM), e.g., [172]. It tries to combine the advantages of FEM and MLM, while alleviating existing drawbacks of MLM. In the XFEM singularities, material discontinuities, high gradients and other non-smooth properties can be described by an extended set of discontinuous basis functions without the need of local remeshing or alignment of the discontinuity (e.g., fractures) to edges or faces of a finite element mesh as usually necessary in standard FEM. However, XFEM is commonly prone to ill-conditioning of the resulting matrix systems, often in a drastic extent, so that standard solution techniques (e.g., preconditioned iterative solvers) are most likely to fail. It is an active field of research to improve the XFEM (e.g., using stable XFEM [17]) for finding more tractable approaches in practical applications.

The SEM represents a combination of the classic spectral method and FEM, e.g., [151, 196, 334]. It can generate solutions of very high accuracy with relatively few terms in the approximate solution, provided that the exact solution is sufficiently smooth (but possibly steep). In contrast to the standard FEM, the unknown coefficients in the approximate solution must not be identified with nodal unknowns. Instead, in SEM formulations the approximate functions are built by Fourier series, Legendre polynomials or Chebyshev polynomials. The main advantage of SEM relies on the exponential convergence property as soon as smooth solutions are involved. For instance, doubling the mesh resolution reduces the numerical error by two orders of magnitude, not by a mere factor of 4 as in standard numerical methods (FEM, FDM, FVM) with second-order algebraic convergence. But, the main drawback of SEM is its inability to handle complex geometries and material discontinuities (even though effort is current to overcome these difficulties, e.g., [413]). It significantly limits the applicability of SEM. Furthermore, SEM has shown insufficently effective for solving linear problems [83].

8.2 Basic Model Equations and Prototypical PDE's

The basic continuum equations of the variable-density flow, mass and heat transport in porous and fractured media have been developed and fully expressed in Chaps. 3 and 4. They have been summarized in Table 3.7 for general variably saturated porous media, in Table 3.9 for fully saturated porous media (groundwater), in Table 3.10 for 2D unconfined aquifers and in Table 3.11 for 2D confined aquifers. Additionally, Tables 4.5–4.7 list the equations for variable-density flow, mass and heat transport of discrete features. A typical set of these coupled governing PDE's can be expressed in the following compact form:

$$\mathcal{L}(\boldsymbol{\phi}) = \begin{cases} \boldsymbol{m} \cdot \dfrac{\partial(\boldsymbol{c} \cdot \boldsymbol{\phi})}{\partial t} + \nabla \cdot (\boldsymbol{f}^a - \boldsymbol{f}^d) - \boldsymbol{b} \qquad\quad = \boldsymbol{0} & \text{divergence form} \\ \boldsymbol{n} \cdot \dfrac{\partial \boldsymbol{\phi}}{\partial t} + \boldsymbol{a} \cdot (\boldsymbol{q} \cdot \nabla \boldsymbol{\phi}) - \nabla \cdot \boldsymbol{f}^d - \boldsymbol{b} + \boldsymbol{d} = \boldsymbol{0} & \text{convective form} \end{cases} \tag{8.1}$$

$$\text{in} \quad \Omega \subset \Re^D, \; t \geq t_0$$

where $\mathcal{L}(\boldsymbol{\phi})$ denotes the PDE system written in terms of the *state variable* $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{x}, t)$. It is expressed on the (physical) domain Ω, with the bounding closure Γ, lying on $D-$dimensional Euclidean space $\Re^D$, and for time t starting at, and proceeding from some initial time t_0. For the solution, appropriate BC's are required on the entirety of Γ and IC's on $\Omega \cup \Gamma$ are necessary as described in Chap. 6. In (8.1) the following vector and matrix definitions are used:

$$\boldsymbol{\phi} = \begin{pmatrix} h \\ C_k \\ T \end{pmatrix}, \; \boldsymbol{f}^a = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{q} C_k \\ \rho c \boldsymbol{q} T \end{pmatrix}, \; \boldsymbol{f}^d = \begin{pmatrix} k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e}) \\ \boldsymbol{D}_k \cdot \nabla C_k \\ \boldsymbol{\Lambda} \cdot \nabla T \end{pmatrix}, \; \boldsymbol{a} = \lceil 0, 1, \rho c \rfloor,$$

$$\boldsymbol{b} = \begin{pmatrix} Q + Q_{\mathrm{EOB}} \\ \tilde{R}_k - \varepsilon s \vartheta_k \Re_k C_k \\ H_e \end{pmatrix}, \; \boldsymbol{d} = \begin{pmatrix} 0 \\ C_k Q \\ \rho c (T - T_0) Q \end{pmatrix}, \; \boldsymbol{m} = \left\lceil s\, S_o + \varepsilon \tfrac{\partial s}{\partial h}, 1, 1 \right\rfloor,$$

$$
\begin{aligned}
\boldsymbol{n} &= \left\lceil s\,S_o + \varepsilon \tfrac{\partial s}{\partial h},\;\; \varepsilon s \acute{\Re}_k,\;\; \varepsilon s \rho c + (1-\varepsilon)\rho^s c^s \right\rfloor, \\
\boldsymbol{c} &= \lceil 1,\;\; \varepsilon s \Re_k,\;\; \varepsilon s \rho c + (1-\varepsilon)\rho^s c^s \rfloor
\end{aligned}
\qquad (8.2)
$$

where $\boldsymbol{f}^a$ and $\boldsymbol{f}^d$ are the advective and dispersive (diffusive) flux tensors, respectively, which are expressed in terms of functions derived from the state variable ϕ. Note that in (8.2) $\lceil \ldots \rfloor$ symbolizes diagonal matrices, cf. (2.22).

The coupled system of PDE's (8.1) has to be solved for ϕ via FEM. Due to its nonlinearity and complexity specific treatments are necessary in dependence on the underlying problem class, e.g., variable-density flow, unsaturated porous media, chemical reaction systems, fracture modeling, heat exchange. It is useful to discuss the FEM solutions for each problem class in a separate manner. However, for introducing the FEM and explaining the principal solution steps it is convenient to start with a simpler PDE written for a scalar state variable ϕ, which is representative for all of the flow and transport processes under consideration. An appropriate prototypical PDE is the following *advection (convection)-dispersion equation* (ADE), incorporating effects of advection, dispersion (diffusion), retardation and decay (as illustrated in Fig. 8.2), written in its *divergence form* as

$$
\begin{aligned}
&\mathcal{L}(\phi) = \frac{\partial(\mathcal{R}\phi)}{\partial t} + \nabla\cdot(\boldsymbol{q}\phi) - \nabla\cdot(\boldsymbol{D}\cdot\nabla\phi) + \vartheta\phi - H - Q_{\phi w} = 0 \\
&\qquad \text{in} \quad \Omega \subset \Re^D,\; t \geq t_0
\end{aligned}
\qquad (8.3)
$$

to be solved for ϕ subject to a set of BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC (see Chap. 6), which typically are

$$
\begin{aligned}
\phi &= \phi_D && \text{on} \quad \Gamma_D \times t[t_0, \infty) \\
(\phi\boldsymbol{q} - \boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n} &= q_N^{\dagger} && \text{on} \quad \Gamma_N \times t[t_0, \infty) \\
(\phi\boldsymbol{q} - \boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n} &= -\Phi^{\dagger}(\phi_C - \phi) && \text{on} \quad \Gamma_C \times t[t_0, \infty) \\
Q_{\phi w} &= -\textstyle\sum_w \phi_w Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty)
\end{aligned}
\qquad (8.4)
$$

and written in its *convective form as*

$$
\begin{aligned}
&\mathcal{L}(\phi) = \acute{\mathcal{R}}\frac{\partial \phi}{\partial t} + \boldsymbol{q}\cdot\nabla\phi - \nabla\cdot(\boldsymbol{D}\cdot\nabla\phi) + (\vartheta + Q)\phi - H - Q_{\phi w} = 0 \\
&\qquad \text{in} \quad \Omega \subset \Re^D,\; t \geq t_0
\end{aligned}
\qquad (8.5)
$$

subject to the Dirichlet, Neumann and Cauchy BC's as well as well-type SPC as

$$
\begin{aligned}
\phi &= \phi_D && \text{on} \quad \Gamma_D \times t[t_0, \infty) \\
-(\boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n} &= q_N && \text{on} \quad \Gamma_N \times t[t_0, \infty) \\
-(\boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n} &= -\Phi(\phi_C - \phi) && \text{on} \quad \Gamma_C \times t[t_0, \infty) \\
Q_{\phi w} &= -\textstyle\sum_w \big(\phi_w - \phi(\boldsymbol{x}_w)\big) Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty)
\end{aligned}
\qquad (8.6)
$$

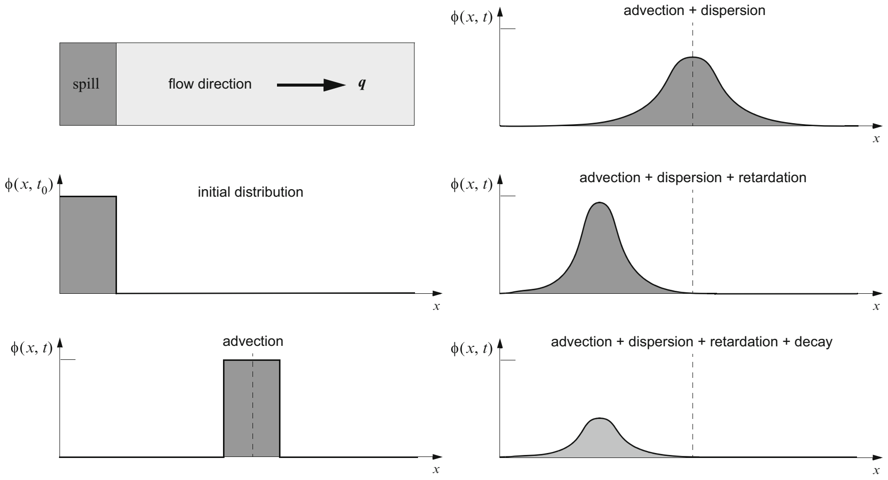


Fig. 8.2 Effects of advection, dispersion, retardation and decay on a scalar transport quantity ϕ: Advection simply translates the quantity by the advective velocity $\boldsymbol{q}$, dispersion spreads the quantity both downstream and upstream and smoothes the fronts, retardation delays the advective transport and reduces effects of dispersion, and decay accounts for disappearance of an amount of the quantity

where the boundary $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_C$ is composed of the three segments, Γ_D, Γ_N and Γ_C, which do not overlap each other: $\Gamma_D \cap \Gamma_N \cap \Gamma_C = \emptyset$. Usually, $\Gamma_D \neq \emptyset$ is required for steady-state problems ($\partial\phi/\partial t = 0$), unless $\Gamma_C \neq \emptyset$. It is assumed that each of the boundary segments can be further subdivided into different portions of the same BC type, e.g., $\Gamma_N = \Gamma_{N_I} \cup \Gamma_{N_O} \cup \ldots$, however, which must not be necessarily connected (Fig. 8.3). The scalar state variable ϕ can stand for the hydraulic head h (or pressure head ψ), for a species concentration C_k or the temperature T in accordance with the corresponding problem class to be solved. In the above Eqs. (8.3)–(8.6), $\boldsymbol{n}$ is the positive outward-directed unit normal to Γ, $\boldsymbol{q}$ is the (at first assuming known) advective flux, $\mathcal{R}$ and $\acute{\mathcal{R}}$ are storage (retardation) coefficients, which are prototypical for the coefficients appearing in (8.2) (note that for an ADE applied to a porous medium they include porosity and saturation), $\boldsymbol{D}$ is a dispersion (diffusion/conduction) tensor, ϑ is a (linear) decay parameter, H is a general source/sink term, Q is a flow supply term (without well-type SPC), $Q_{\phi w}$ is the singular well sink/source function with given well pumping rate $Q_w(t)$ and known ϕ_w at well w of location $\boldsymbol{x}_w$, ϕ_D is the prescribed value of ϕ on the Dirichlet boundary segment Γ_D, $q_N^\dagger$ and q_N are the prescribed fluxes on the Neumann boundary segment Γ_N for the divergence and the convective form of the ADE, respectively, and ϕ_C is a known value of ϕ on the Cauchy boundary segment Γ_C associated with the transfer coefficients $\Phi^\dagger$ and Φ related to the divergence and convective form of the governing ADE, respectively. Note that $q_N^\dagger$ and q_N as well as ϕ_C associated with $\Phi^\dagger$ or Φ have different meaning in the divergence form (8.3) and the convective form (8.6) since in the divergence form the boundary flux consists of

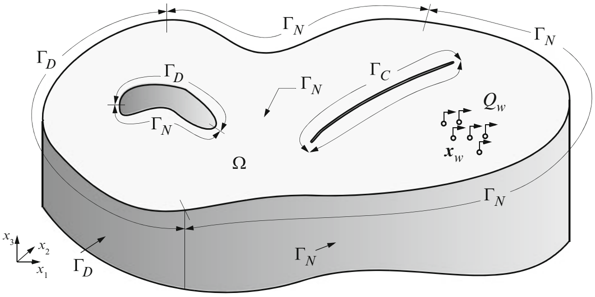


Fig. 8.3 Domain $\Omega \subset \Re^3$ and boundary sections of Dirichlet type $\Gamma_D \subset \Gamma$, Neumann type $\Gamma_N \subset \Gamma$ and Cauchy type $\Gamma_C \subset \Gamma$ as well as SPC's Q_w for wells at $\boldsymbol{x}_w \in \Omega$

the total (advective plus dispersive) flux while in the convective form the boundary flux implies only a dispersive flux. However, as has been shown in Sect. 6.3.2 the Neumann-type BC of the divergence form is equivalent to the Cauchy-type BC of (8.6) if using for the convective form, cf. (6.21) and (6.28),

$$-(\boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n} = \boldsymbol{q}\cdot\boldsymbol{n}(\phi_C-\phi) \quad \text{on} \quad \Gamma_C \times t[t_0,\infty) \tag{8.7}$$

Finally, the statement of the PDE problem (8.3) or (8.5) has to be completed by specifying an IC in the form:

$$\phi(\boldsymbol{x},t_0) = \phi_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\Omega} \tag{8.8}$$

where ϕ_0 is a given function of ϕ at position $\boldsymbol{x}$ and initial time t_0 with $\bar{\Omega} = \Omega \cup \Gamma$.

8.3 Mathematical Classification of PDE's

The governing partial differential equations (PDE's), such as summarized in Sect. 8.2, can be mathematically classified into three categories: (1) elliptic, (2) parabolic and (3) hyperbolic. Most of the equations are 2nd-order PDE's. To classify the PDE's several procedures are available, where most common is the discriminant evaluation [486]: Let us consider the PDE of the form in a 2D domain $\boldsymbol{x}^T = (x\ y) \in \Re^2$

$$\mathcal{L}(\phi) = A\frac{\partial^2\phi}{\partial x^2} + B\frac{\partial^2\phi}{\partial x\partial y} + C\frac{\partial^2\phi}{\partial y^2} + D\frac{\partial\phi}{\partial x} + E\frac{\partial\phi}{\partial y} + F\phi + G = 0 \tag{8.9}$$

where the coefficients A, B, C, D, E, F and G are constants or may be functions of both independent and/or dependent variables. Then, the three categories of PDE can be distinguished according to:

$$\begin{aligned}&\text{elliptic PDE:} && B^2-4AC<0,\\ &\text{parabolic PDE:} && B^2-4AC=0,\\ &\text{hyperbolic PDE:} && B^2-4AC>0\end{aligned} \tag{8.10}$$

It is apparent that the classification depends only on the highest-order derivatives in each independent variable. We note that the coefficients A to G of (8.9) can also vary as functions of x, y, ϕ, $\partial\phi/\partial x$ or $\partial\phi/\partial y$ and (8.10) can still be used if A, B and C are given a local interpretation. This implies that an equation can belong to one classification in one part of the domain and another classification in another part of the domain. Typical examples of PDE classifications are given as follows:

(a) Elliptic equation

$$\begin{aligned}&-K_x\frac{\partial^2\phi}{\partial x^2}-K_y\frac{\partial^2\phi}{\partial y^2}=0 \quad (K_x>0,\;\; K_y>0)\\ &A=-K_x,\; B=0,\; C=-K_y\\ &B^2-4AC=-4K_xK_y<0\end{aligned} \tag{8.11}$$

(b) Parabolic equation

$$\begin{aligned}&\frac{\partial\phi}{\partial t}+v\frac{\partial\phi}{\partial x}-K_x\frac{\partial^2\phi}{\partial x^2}=0 \quad (v>0,\; K_x>0)\\ &A=-K_x,\; B=0,\; C=0\\ &B^2-4AC=0\end{aligned} \tag{8.12}$$

(c) Hyperbolic equation

$$\begin{aligned}&\frac{\partial\phi}{\partial t}+v\frac{\partial\phi}{\partial x}=0 \quad (v>0)\\ &\text{differentiating with respect to } x \text{ and } t\text{:}\\ &\frac{\partial^2\phi}{\partial t\partial x}+v\frac{\partial^2\phi}{\partial x^2}=0, \qquad \frac{\partial^2\phi}{\partial t^2}+v\frac{\partial^2\phi}{\partial t\partial x}=0\\ &\text{and combining:}\\ &\frac{\partial^2\phi}{\partial t^2}-v^2\frac{\partial^2\phi}{\partial x^2}=0\\ &A=1,\; B=0,\; C=-v^2\\ &B^2-4AC=4v^2>0\end{aligned} \tag{8.13}$$

To generalize the PDE classification to more variables, a common way is to classify the PDE's via the 2nd-order differential operator defined as

$$\mathcal{D}(\phi)=\nabla\cdot(\boldsymbol{D}\cdot\nabla\phi) \tag{8.14}$$

where $\boldsymbol{D}$ is a symmetric, positive-definite tensor. Then, the three categories of PDE in the $D-$dimensional space $\Re^D$ $(D = 1, 2, 3)$ are distinguished as follows:

$$\begin{aligned} &\text{elliptic PDE:} && -\mathcal{D}(\phi) - F\phi + G = 0, \\ &\text{parabolic PDE:} && a\frac{\partial \phi}{\partial t} + \boldsymbol{v}\cdot\nabla\phi - \mathcal{D}(\phi) - F\phi + G = 0, \\ &\text{hyperbolic PDE:} && a\frac{\partial \phi}{\partial t} + \boldsymbol{v}\cdot\nabla\phi + F\phi - G = 0 \end{aligned} \tag{8.15}$$

where a, F and G are coefficients and $\boldsymbol{v}$ represents a flux vector.

The classification of PDE's can be associated with the smoothness of the solution ϕ. Elliptic PDE's produce solutions that are smooth (up to the smoothness of coefficients) even if BC's are not smooth. On the other hand, parabolic PDE's will cause the smoothness of solutions to increase with growing time and reducing influences by first-order derivatives, while hyperbolic PDE's preserve lack of smoothness.

8.4 Methods of Approximation

8.4.1 Approximate Solution

The sought approximation of the basic PDE's (8.3) and (8.5) with their BC's (8.4) and (8.6), respectively, starts with expressing a suitable approximate functional form for the solution ϕ. The usual form is

$$\phi(\boldsymbol{x}, t) \approx \hat{\phi}(\boldsymbol{x}, t) = \sum_j N_j(\boldsymbol{x})\,\phi_j(t) \tag{8.16}$$

where $\hat{\phi}$ is the approximate solution, N_j represent a set of given *basis functions* (or trial or interpolation functions) and ϕ_j are a set of unknown coefficients (at the *nodes* of interpolation) to be determined. In the functional expression (8.16) the spatial and temporal variables are separated. This variable separation procedure is termed as *Kantorovich (semidiscrete) method* [149, 300, 377] and allows the discretization first in space followed by a time marching procedure for the temporal discretization, which is the usual practice in numerical analysis leading to efficient computational schemes, although alternative of (8.16) exists.[1]

[1] A continuous space-time approximation can be expressed in the form

$$\phi(\boldsymbol{x}, t) \approx \hat{\phi}(\boldsymbol{x}, t) = \sum_j N_j(\boldsymbol{x}, t)\,\phi_j$$

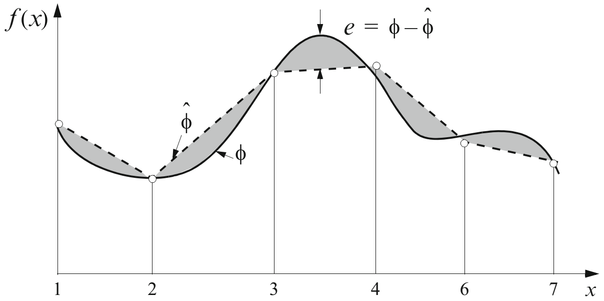


Fig. 8.4 Approximating $\hat{\phi}(x)$ (*dashed line*) of state function ϕ (*solid line*) creating an error $e = \phi - \hat{\phi}$ (*shaded area*)

It is to be noted that the sought approximate solution $\hat{\phi}$ is a function distributed over the entire domain Ω of $\mathcal{L}(\phi)$ and its boundary Γ; hence, it is a *global* function. Examples of the corresponding basis functions include polynomials (e.g., Lagrangian, Hermite or Chebyshev polynomials) or trigonometric functions (e.g., Fourier series). Approximating the solution to (8.3)–(8.6) with the series expression (8.16), an *error* will generally occur defined as the difference between the exact solution ϕ and the approximate solution $\hat{\phi}$:

$$e = \phi - \hat{\phi} \tag{8.17}$$

The situation is sketched in Fig. 8.4, where the exact solution ϕ shown as solid line is approximated by a piecewise continuous linear interpolation $\hat{\phi}$ between selected locations at nodes $j = 1, 2, \ldots$ depicted as dashed line. The difference between ϕ and $\hat{\phi}$ represents the error e of the solution illustrated by the shaded area in Fig. 8.4. It indicates that the error is in general a function of space (and time). The error can also be measured only at the discrete locations of nodes $j = 1, 2, \ldots$, providing a vector $\boldsymbol{e}$ of pointwise errors:

$$\boldsymbol{e} = e_j = \phi_j - \hat{\phi}_j \quad (j = 1, 2, \ldots) \tag{8.18}$$

The goal is now to make the error e as small as possible, and hence minimize the difference between ϕ and $\hat{\phi}$. Since the exact solution ϕ is generally unknown and e is variable in space and time, the minimization of the error e requires a general approach to be described in the following.

where the basis functions $\boldsymbol{x}, t$ have to be prescribed both in space $\boldsymbol{x}$ and time t. It requires a finite element in space-time and increases the computational dimension, e.g., a transient 3D problem needs a 4D trial space.

8.4.2 Definition of Errors and Related Functional Spaces

Defining in the $D-$dimensional Euclidean $\Re^D$ space with

$$D^s e = \frac{\partial^{|s|} e(\boldsymbol{x})}{\partial x_1^{s_1} \partial x_2^{s_2} \ldots \partial x_D^{s_D}} = \left(\frac{\partial^{s_1}}{\partial x_1^{s_1}}\right)\left(\frac{\partial^{s_2}}{\partial x_2^{s_2}}\right) \cdots \left(\frac{\partial^{s_D}}{\partial x_D^{s_D}}\right) e(\boldsymbol{x}) \tag{8.19}$$

the generalized partial derivatives up to and including of the $2m$th order appearing in the governing PDE's ($m = 0, 1, \ldots$), where s is a multi-index $s = (s_1, s_2, \ldots)$ with $|s| = \sum_{i=1}^{D} s_i$ and $s_i = 0, 1, \ldots$, the following error norms are meaningful in the further analysis.

8.4.2.1 Sobolev Space $W_p^m(\Omega)$ Norm Error

The Sobolev space norm error is defined as

$$\|e\|_{W_p^m(\Omega)} = \left\{ \int_\Omega \left[|e|^p + \sum_{s=1}^{m} |D^s e|^p \right] d\Omega \right\}^{\frac{1}{p}} \tag{8.20}$$

where m denotes the highest order of the derivatives of the $2m$th governing PDE and p represents the power to which the derivatives are raised. Note that for a 2nd order PDE $m = 1$. The Sobolev space $W_p^m(\Omega)$ is defined as the functional space which includes all p integrable functions $(1 \leq p \leq +\infty)$ with p integrable derivatives of mth order. Hence, $W_p^m(\Omega)$ is a collection of functions on Ω which are endowed with the associated norm (8.20), where any function $\phi \in W_p^m(\Omega)$ is m times differentiable and pth-order integrable on Ω.

8.4.2.2 Hilbert Space $H^m(\Omega)$ Norm Error

The Hilbert space $H^m(\Omega)$ corresponds to the Sobolev space $W_p^m(\Omega)$ with p equal to 2, i.e., $H^m(\Omega) = W_2^m(\Omega)$. Thus,

$$\|e\|_{H^m(\Omega)} = \|e\|_{W_2^m(\Omega)} = \left\{ \int_\Omega \left[e^2 + \sum_{s=1}^{m} (D^s e)^2 \right] d\Omega \right\}^{\frac{1}{2}} \tag{8.21}$$

As seen the Hilbert space $H^m(\Omega)$ is a functional space with square integrable functions and square integrable derivatives of mth order. Any function $\phi \in H^m(\Omega)$ is m times differentiable and square integrable on Ω.

8.4.2.3 Energy Norm Error

The energy norm error $\|e\|_E$ is a special case of the Hilbert space norm $H^m(\Omega)$ in the $2m$the PDE. For a 2nd-order PDE ($m = 1$) it reads

$$\|e\|_E = \|e\|_{H^1(\Omega)} = \|e\|_{W_2^1(\Omega)} = \left\{ \int_\Omega \left[e^2 + \left(\frac{\partial e}{\partial x_1} \right)^2 + \left(\frac{\partial e}{\partial x_2} \right)^2 + \ldots \right] d\Omega \right\}^{\frac{1}{2}} \tag{8.22}$$

The Hilbert space $H^1(\Omega)$ is a functional space with square integrable functions and square integrable derivatives of 1st order. Any function $\phi \in H^1(\Omega)$ is once differentiable and square integrable on Ω. A (smaller) Hilbert subspace $H_0^1(\Omega)$ can be defined for functions ϕ which are zero on the boundary Γ of the domain Ω at the same time, i.e., $\phi|_\Gamma = 0$. Then, the Hilbert subspace $H_0^1(\Omega)$ reads

$$H_0^1(\Omega) \equiv \left\{ \phi \in H^1(\Omega) : \ \phi = 0 \ \text{ on } \ \Gamma \right\} \tag{8.23}$$

so that any function $\phi \in H_0^1(\Omega)$ is once differentiable, square integrable on Ω and zero on Γ.

8.4.2.4 $L_p(\Omega)$–Norm (Banach Space Norm) Error

The Banach space $L_p(\Omega)$ is defined as the complete normed linear space such that

$$\|e\|_{L_p(\Omega)} = \left(\int_\Omega |e|^p d\Omega \right)^{\frac{1}{p}} \tag{8.24}$$

Using $p = 2$ we obtain the $L_2(\Omega)$ space, which is equivalent to the Hilbert space $H^0(\Omega)$ with $m = 0$:

$$\|e\|_{L_2(\Omega)} = \|e\|_{H^0(\Omega)} = \|e\|_{W_2^0(\Omega)} = \left(\int_\Omega e^2 d\Omega \right)^{\frac{1}{2}} \tag{8.25}$$

The $L_2(\Omega)$ space is a functional space with square integrable functions so that any function $\phi \in L_2(\Omega)$ must be square integrable on Ω. The L_2 norm is one of the most widely used error norm. Another useful error norm is the maximum error norm $L_\infty(\Omega)$ given for $p = \infty$:

$$\|e\|_{L_\infty(\Omega)} = \max_j |e_j| \tag{8.26}$$

where e_j is the discrete error at location j. More seldom used in practice is the $L_1(\Omega)$ error norm:

$$\|e\|_{L_1(\Omega)} = \int_\Omega |e| \, d\Omega \tag{8.27}$$

8.4.2.5 Root Mean Square (RMS) and Other Pointwise Error Norms

The *RMS error norm* represents a pointwise $L_2(\Omega)$ space norm, which can be expressed in different forms. Most useful is the normalized RMS error norm defined as

$$\|e\|_{\mathrm{RMS}} = \left[\frac{1}{N_{\mathrm{P}}}\left(\frac{1}{\hat{\phi}^2_{\max}}\sum_{j=1}^{N_{\mathrm{P}}} e_j^2\right)\right]^{\frac{1}{2}} \tag{8.28}$$

where N_{P} is the number of components of the error vector $\boldsymbol{e}$ and $\hat{\phi}_{\max}$ is the maximum value of the approximate solution $\hat{\phi}$ to normalize the e_j components.

If focusing on the maximum error occurring in the approximate discrete solution, the *normalized maximum error norm* can be useful and is defined as

$$\|e\|_{L_\infty} = \frac{1}{\hat{\phi}_{\max}} \max_j |e_j| \tag{8.29}$$

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

As an alternative to the L_2 RMS norm, the normalized L_1 error norm can be applied:

$$\|e\|_{L_1} = \frac{1}{N_{\mathrm{P}}\,\hat{\phi}_{\max}} \sum_{j=1}^{N_{\mathrm{P}}} |e_j| \tag{8.30}$$

However, it should not be the first choice and the RMS norm is commonly more appropriate.

8.4.3 Method of Weighted Residuals (MWR)

There are two fundamental theories of constructing approximate solutions to the governing PDE’s:

1. The classic *Rayleigh-Ritz method* [377,590], which is based on finding solutions via an equivalent variational problem. By extremization of the related variational functional (condition of stationarity) useful approximate solutions can be obtained. However, natural variational functionals only exist for *self-adjoint*[2]

[2] Let $\mathcal{L}$ be a differential operator of a PDE defined in Ω and let ϕ and ψ be two functions in the field of definition of $\mathcal{L}$. The operator $\mathcal{L}$ is said to be self-adjoint if identical to its own adjoint operator $\mathcal{L}^*$, i.e., $\mathcal{L} = \mathcal{L}^*$, which must result from the integral statement

differential operator $\mathcal{L}$ of the governing PDF. A self-adjoint PDE is given for a symmetric equation (containing no advective terms). However, ADE in the form of (8.3) or (8.5) possesses an unsymmetric non-self-adjoint differential operator for which a natural variational functionals cannot be found.[3]

2. The *method of weighted residuals* (MWR) [163] provides the most generality in applications and will be preferred usually. It can be applied to all type of PDE and systems of PDE's, even to those which cannot be cast in variational form. The following finite element approach will be exclusively based on MWR.

It is obvious that the approximate solution $\hat{\phi}$ of (8.16) is not likely to satisfy exactly the governing PDE

$$\mathcal{L}(\phi) = 0 \tag{8.31}$$

in form of (8.3) or (8.5). Substituting $\hat{\phi}$ in (8.31) yields a PDE for the error $e = \phi - \hat{\phi}$, (8.17), written as

$$\mathcal{L}(e) = \mathcal{L}(\phi) - \mathcal{L}(\hat{\phi}) = -\mathcal{L}(\hat{\phi}) \neq 0 \tag{8.32}$$

or

$$\mathcal{L}(\hat{\phi}) = R, \quad R = -\mathcal{L}(e) \neq 0 \tag{8.33}$$

where $R = R(\boldsymbol{x}, t)$ is the *residual*, which is a measure of the *induced error* arising from the used approximation. It is commonly impossible and not reasonable to try to force R to be zero everywhere in Ω and on Γ (it would meet the exact solution).

$$\int_\Omega \mathcal{L}(\phi)\psi \, d\Omega = \int_\Omega \phi \mathcal{L}^*(\psi) \, d\Omega + \text{boundary integral terms}$$

[3] For instance, the non-self-adjoint ADE in form of (8.5) can be transformed to a self-adjoint problem by introducing the new operator [129, 132, 218, 427]

$$\bar{\mathcal{L}} = \varphi \, \mathcal{L}$$

where the function $\varphi = \varphi(\boldsymbol{x})$ is chosen by

$$\varphi = \exp(\beta), \quad \beta = -\frac{\boldsymbol{q} \cdot \boldsymbol{x}}{\|\boldsymbol{D}\|}$$

assuming a dispersion tensor $\boldsymbol{D}$ with $D_{ij} = 0$ for $i \neq j$. It yields the following variational functional

$$\mathcal{I} = \int_\Omega \Big[\tfrac{1}{2} \nabla\phi \cdot (\boldsymbol{D} \cdot \nabla\phi) + \Big(\acute{\mathcal{R}} \frac{\partial \phi}{\partial t} + \frac{\vartheta + Q}{2} \phi - H - Q_{\phi w}\Big)\phi\Big] \exp(\beta) d\Omega - \int_\Gamma (\boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n}\phi \exp(\beta) d\Gamma$$

to be extremized. However, its application is clearly restricted because values of $\exp(\beta)$ can become very large (small) for advection-dominated processes and the variational functional terms overflow (underflow) in practical computations [129, 485].

Instead, the pragmatic approach is to require the residual R to vanish in an overall integrated sense. The corresponding mathematical statement is that R must be *orthogonal*[4] to an arbitrary *weighting* (or test) *function* $w(\boldsymbol{x}, t)$, i.e.,

$$\int_{\Omega} w(\boldsymbol{x}, t)\, R\, d\Omega = 0, \quad \text{for all } w(\boldsymbol{x}, t) \tag{8.34}$$

The expression (8.34) is the core of MWR [163], which minimizes the residual R as a weighted average over the domain Ω. This form is quite general and the arbitrariness in $w(\boldsymbol{x}, t)$ provides theoretical generality for various numerical approaches. For specifying appropriate weighting functions $w(\boldsymbol{x}, t)$ it is assumed that its interpolation, using any suitable polynomial basis, can be made sufficiently precise:

$$w(\boldsymbol{x}, t) \approx \hat{w}(\boldsymbol{x}, t) = \sum_i w_i(\boldsymbol{x}) W_i(t) \tag{8.35}$$

where $w_i(\boldsymbol{x})$ is the set of interpolation polynomials and $W_i(t)$ is the corresponding set of known coefficients at the nodes of interpolation. The coefficients $W_i(t)$,

[4] We know from (2.26) when two vectors in space are at right angles, their dot product is zero and the vectors are orthogonal. While vectors have only a limited number of entries, any real-valued function $f(\boldsymbol{x})$ is characterized by infinite number of points within its domain of definition Ω. It is obvious to consider two functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ to be orthogonal, if the product $f(\boldsymbol{x})g(\boldsymbol{x})$ 'summed' over all $\boldsymbol{x}$ within the domain Ω results zero. Since the amount of $\boldsymbol{x}$ covers infinite real numbers, the product $f(\boldsymbol{x})g(\boldsymbol{x})$ has to be integrated. Hence, the analogy for the dot product is the *inner product* given by

$$(f, g) = \int_{\Omega} f(\boldsymbol{x}) g(\boldsymbol{x}) d\Omega$$

Then, the two functions are orthogonal if $(f, g) = 0$. It is evident that functions if defined in the $L_2(\Omega)$ space, cf. (8.25), e.g.,

$$\|f\| = \Big(\int_{\Omega} f(\boldsymbol{x})^2 d\Omega\Big)^{\frac{1}{2}} < \infty, \quad \|g\| = \Big(\int_{\Omega} g(\boldsymbol{x})^2 d\Omega\Big)^{\frac{1}{2}} < \infty$$

can be treated if they were vectors, where the *Schwarz's inequality* holds

$$(f, g) \leq \|f\|\|g\|$$

or

$$|(f, g)|^2 \leq (f, f)(g, g)$$

The $L_2(\Omega)-$norm corresponds to a measure of the size of a function, which is in direct analogy with the vector norm (2.11). The Schwarz's inequality ensures that the expression

$$\cos\theta = \frac{(f, g)}{\|f\|\|g\|}$$

yields well-defined angles θ in space similar to the scalar product of vectors (2.24).

Table 8.1 Suitable choices of weighting functions $w_i(\boldsymbol{x})$, $(i = 1, 2, \ldots, N_{\text{EQ}})$ and their resulting numerical methods

$w_i(\boldsymbol{x})$	WS (8.36)[a]	Method	Remark
$\delta(\boldsymbol{x} - \boldsymbol{x}_i)$	$\int_\Omega \delta(\boldsymbol{x} - \boldsymbol{x}_i)\mathcal{L}(\hat{\phi})d\Omega = 0$	Point collocation	FDM
$\begin{cases} 1 & \text{for } \boldsymbol{x}_i \in \Omega^e \\ 0 & \text{for } \boldsymbol{x}_i \notin \Omega^e \end{cases}$	$\int_\Omega \mathcal{L}(\hat{\phi})d\Omega = 0$	Subdomain collocation	FVM
N_i	$\int_\Omega N_i(\boldsymbol{x})\mathcal{L}(\hat{\phi})d\Omega = 0$	Galerkin (Bubnov-Galerkin)	GFEM (standard FEM)
$N_i + \tilde{F}_i(\boldsymbol{x})$	$\int_\Omega \big(N_i(\boldsymbol{x}) + \tilde{F}_i(\boldsymbol{x})\big)\mathcal{L}(\hat{\phi})d\Omega = 0$	Petrov-Galerkin	PGFEM (upwind)
$\partial R/\partial \phi_i$	$\int_\Omega \mathcal{L}\big(N_i(\boldsymbol{x})\big)\mathcal{L}(\hat{\phi})d\Omega = 0$	Least square Galerkin	LSGFEM (PGLS)

[a] $\mathcal{L}(\hat{\phi}) = R = \mathcal{L}\Big(\sum_{j=1}^{N_{\text{P}}} N_j(\boldsymbol{x})\phi_j(t)\Big)$

however, quantify the specific weighting function w. To remove this dependence on a specific w, the weak statement (8.35) is extremized with respect to the parametric set $W_i(t)$. Thus, the following weak statement (WS) for minimizing the residual error $R = \mathcal{L}(\hat{\phi})$ in any selected approximate solution $\hat{\phi}$ (8.16) results

$$\begin{aligned} \text{WS} &= \frac{\partial}{\partial W_i} \int_\Omega \hat{w}(\boldsymbol{x}, t)\mathcal{L}(\hat{\phi})d\Omega = 0 \\ &= \int_\Omega w_i(\boldsymbol{x})\mathcal{L}\Big(\sum_{j=1}^{N_{\text{P}}} N_j(\boldsymbol{x})\phi_j(t)\Big)d\Omega = 0 \quad \text{for } (1 \leq i \leq N_{\text{EQ}}) \end{aligned} \tag{8.36}$$

where $(i = 1, 2, \ldots, N_{\text{EQ}})$ is chosen to produce exactly the correct number of equations required to determine the N_{EQ} unknown coefficients $\phi_j(t)$ at any time t. We note that

$$N_{\text{EQ}} = N_{\text{P}}\, N_{\text{DOF}} \tag{8.37}$$

where N_{P} is the number of chosen nodes and N_{DOF} is the number of *degrees of freedom*. For example, $N_{\text{DOF}} = 1$ for scalar equations of ϕ in the form of (8.3) or (8.5) and $N_{\text{DOF}} = N + 2$ for the vectorial variable $\phi = (h\ \ C_k\ \ T)^T$ $(k = 1, 2, \ldots, N)$ appearing in (8.2). Having the weak statement expressed in the form of (8.36) it remains to identify the two sets of known functions $w_i(\boldsymbol{x})$ and $N_j(\boldsymbol{x})$ spanning the domain $\Omega \subset \Re^D$. Usually, both the basis function set $N_j(\boldsymbol{x})$ and the weighting function set $w_i(\boldsymbol{x})$ are defined as interpolation polynomials, with a typical selection as Lagrange polynomials. Depending on the choice of the weighting functions $w_i(\boldsymbol{x})$ various alternative (and familiar) methods can be generated. The most important methods are summarized in Table 8.1.

Viewing Table 8.1, we can recognize classic numerical techniques as special cases of MWR. In the *point collocation* approach a set of points $\boldsymbol{x}_i$ is specified in the solution domain Ω and Dirac delta functions are chosen as weighting functions.

It produces a discrete approximation referred to as a stencil, common in finite-difference schemes. The choices of polynomials $N_j(\boldsymbol{x})$ determine finally the accuracy of the finite-difference approximation. In the *subdomain collocation* approach the solution domain is subdivided into a number of subdomains $\Omega = \cup\Omega^e$ and weighting functions are unity for all i if $\boldsymbol{x}_i \in \Omega^e$ and zero otherwise. It leads to finite volume approximations. As w_i are constant in each of the respective subdomains, any integration by part reduces to boundary integrals. First-order operations are obvious and give normal fluxes through the discretized subdomain boundaries. However, with derivatives higher than first-order, FVM approaches require specific treatment such as TPFA or MPFA schemes [412].

A suitable option for the set of weighting functions $w_i(\boldsymbol{x})$ is to require it be identical to the set of basis functions $N_i(\boldsymbol{x})$ by each term i: $w_i = N_i$. It means, the test functions are represented by a linear combination of the *same* basis functions as used to approximate the solution. This is known as the *Galerkin criterion* named after B.G. Galerkin [181] who originally introduced it for (non-discrete) structural formulations. This Galerkin method leads to the standard finite-element approximation, called as Galerkin-FEM or GFEM (sometimes termed as Bubnov-Galerkin method [590] to differ from the modified Petrov-Galerkin method). It is important to note that the Galerkin-based WS enforces the residual error R be orthogonal to every member of the basis functions, which provides an *optimal* approximation expressed by *Céa's lemma* [84, 193, 555] written in the form:

$$\|e\|_{E,\,\mathrm{G}} \leq \|e\|_{E,\,\mathrm{O}} \tag{8.38}$$

where $\|e\|_{E,\,\mathrm{G}}$ and $\|e\|_{E,\,\mathrm{O}}$ are the energy (Hilbert space) norm errors (8.22) produced by the Galerkin method and by any other approximation method, respectively. For elliptic boundary value problems the optimality (8.38) is explicitly shown in Appendix F. Extensions to GFEM are given in the so-called Petrov-Galerkin method, where the weighting functions differ from the basis functions. It allows the foundation of stabilized numerical techniques which are appropriate for solving advection-dominated transport problems.

In *least squares* (LS) the set of weighting functions is constructed via the PDE operation. The resulting schemes can provide better convergence properties. Furthermore, it can be exploited to derive stabilized methods for ADE with dominant advection. An advantageous and attractive feature of the LS method is that a non-self-adjoint (1st-order differential) operator of PDE is converted into a self-adjoint 2nd-order problem, which provides *symmetry* in the approximate equation system. The Galerkin choice $w_i(\boldsymbol{x}) = N_i(\boldsymbol{x})$ is also optimal for LS approximations.

In the following Galerkin WS will be taken as the base finite-element weak statement. Extensions will be given for the Petrov-Galerkin and least square FEM to derive artificial diffusion stabilization mechanisms of upwind schemes applied to ADE.

8.5 Weak Forms

For the following finite element analysis any governing PDE $\mathcal{L}(\phi) = 0$ (with its BC's and IC's) has to be recast into its weak form (or weak statement) according to (8.34)

$$\int_{\Omega} w(\boldsymbol{x}, t)\, \mathcal{L}(\phi)\, d\Omega = 0, \quad \forall w(\boldsymbol{x}, t) \tag{8.39}$$

where $w(\boldsymbol{x}, t)$ is an arbitrary weighting function. It is important to note the difference between the original PDE formulation and the weak form from the mathematical point of view. While the classic statement of the initial boundary value problem is in general unique and unambiguous, there is usually no unique weak statement of the same problem because there are alternative choices for w and optional formulations for BC's. Each weak form, however, has usually a unique solution. Some weak statements are more useful than others and it is important to find the most appropriate weak form. In this sense, a weak form represents a formulation *equivalent* to the governing PDE. The weak form incorporates the BC's.

8.5.1 Divergence Form of ADE

The weak form (8.39) in application to the ADE (8.3) yields

$$\int_{\Omega} w \frac{\partial(\mathcal{R}\phi)}{\partial t} d\Omega + \int_{\Omega} w\nabla \cdot (\boldsymbol{q}\phi) d\Omega - \int_{\Omega} w\nabla \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega + \int_{\Omega} w(\vartheta\phi - H - Q_{\phi w}) d\Omega = 0 \tag{8.40}$$

which is satisfied for any weighting function $w = w(\boldsymbol{x}, t)$. In the formulation of (8.40) w need not to be differentiable and it is sufficient to require that w is only square integrable: $\forall w \in L_2(\Omega)$.

However, let us restrict the class of weighting functions to those, which are at least once-differentiable, i.e., $\forall w \in H^1(\Omega)$. The restriction on w permits to invoke the following identity via partial integration applied to the 1st-order advective term

$$\int_{\Omega} \nabla \cdot (w\boldsymbol{q}\phi) d\Omega = \int_{\Omega} w\nabla \cdot (\boldsymbol{q}\phi) d\Omega + \int_{\Omega} \phi\boldsymbol{q} \cdot \nabla w d\Omega \tag{8.41}$$

and to the 2nd-order dispersion term

$$\int_{\Omega} \nabla \cdot [w(\boldsymbol{D} \cdot \nabla\phi)] d\Omega = \int_{\Omega} w\nabla \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega \tag{8.42}$$

Now, let us apply the Gauss's integral theorem (2.77) to the LHS's of (8.41) and (8.42) to obtain

$$\begin{aligned}\int_{\Omega} \nabla \cdot (w\boldsymbol{q}\phi) d\Omega &= \int_{\Gamma} w\phi\boldsymbol{q} \cdot \boldsymbol{n}\, d\Gamma \\ \int_{\Omega} \nabla \cdot [w(\boldsymbol{D} \cdot \nabla\phi)] d\Omega &= \int_{\Gamma} w(\boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n}\, d\Gamma\end{aligned} \tag{8.43}$$

and to find for (8.41)

$$\int_{\Omega} w\nabla \cdot (\boldsymbol{q}\phi) d\Omega = \int_{\Gamma} w\phi\boldsymbol{q} \cdot \boldsymbol{n}\, d\Gamma - \int_{\Omega} \phi\boldsymbol{q} \cdot \nabla w d\Omega \tag{8.44}$$

and for (8.42)

$$\int_{\Omega} w\nabla \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega = \int_{\Gamma} w(\boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n}\, d\Gamma - \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega \tag{8.45}$$

Inserting (8.44) and (8.45) into (8.40), the weak form becomes

$$\begin{aligned}&\int_{\Omega} w\frac{\partial(\mathcal{R}\phi)}{\partial t} d\Omega - \int_{\Omega} \phi\boldsymbol{q} \cdot \nabla w d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega + \\ &\int_{\Omega} w(\vartheta\phi - H - Q_{\phi w}) d\Omega + \int_{\Gamma} w(\phi\boldsymbol{q} - \boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n}\, d\Gamma = 0, \quad \forall w \in H^1(\Omega)\end{aligned} \tag{8.46}$$

Recalling that the boundary is composed of three segments $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_C$ imposed by the Dirichlet, Neumann and Cauchy-type BC's, we can separate the boundary integral of (8.46) into these three parts and invoke the BC's of (8.4) to obtain

$$\begin{aligned}&\int_{\Omega} w\frac{\partial(\mathcal{R}\phi)}{\partial t} d\Omega - \int_{\Omega} \phi\boldsymbol{q} \cdot \nabla w d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega + \\ &\int_{\Omega} w(\vartheta\phi - H - Q_{\phi w}) d\Omega + \int_{\Gamma_D} w(\phi\boldsymbol{q} - \boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n}\, d\Gamma + \\ &\int_{\Gamma_N} wq_N^{\dagger} d\Gamma - \int_{\Gamma_C} w\Phi^{\dagger}(\phi_C - \phi) d\Gamma = 0, \quad \forall w \in H^1(\Omega)\end{aligned} \tag{8.47}$$

Now, we have to further restrict the class of test functions w to those that vanish on the Dirichlet boundary segment Γ_D, i.e., we require $w = 0$ on Γ_D. This class of functions belongs to the H_0^1 functional space (8.23). Using this restriction of $\forall w \in H_0^1$, the final weak form for the divergence form of ADE (8.3) with its BC's (8.4) results

$$\int_{\Omega} w \frac{\partial(\mathcal{R}\phi)}{\partial t} d\Omega - \int_{\Omega} \phi \boldsymbol{q} \cdot \nabla w d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega +$$
$$\int_{\Omega} w(\vartheta\phi - H) d\Omega + \sum_{w} w(\boldsymbol{x}_w)\phi_w Q_w(t) + \int_{\Gamma_N} w q_N^{\dagger} d\Gamma -$$
$$\int_{\Gamma_C} w \Phi^{\dagger}(\phi_C - \phi) d\Gamma = 0, \quad \forall w \in H_0^1(\Omega) \tag{8.48}$$

which has to be solved for $\phi \approx \hat{\phi}$. We recognize from (8.48) that the sought solution must also only be once differentiable, i.e., $\phi \approx \hat{\phi} \in H^1(\Omega)$. Note that in (8.48) the well-type SPC (8.4) has been inserted, where we made use of the integral over the SPC singularity, which simplifies

$$\int_{\Omega} w Q_{\phi w} d\Omega = -\int_{\Omega} w\big(\sum_{w} \phi_w Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w)\big) d\Omega = -\sum_{w} w(\boldsymbol{x}_w)\phi_w Q_w(t) \tag{8.49}$$

This SPC realization in the weak form implies that the entire amount of the sink/source Q_w of a well w fully pertains to the equation at the given point $\boldsymbol{x}_w$. In a finite element approximation it will be attained by enforcing that each well coincides with a node of the spatial discretization.

8.5.2 Convective Form of ADE

The weak form for the ADE (8.5) with its BC's (8.6) can be derived in a similar way as done in Sect. 8.5.1 for the divergence form. The weak statement (8.39) applied to (8.5) yields

$$\int_{\Omega} w \acute{\mathcal{R}} \frac{\partial \phi}{\partial t} d\Omega + \int_{\Omega} w \boldsymbol{q} \cdot \nabla\phi d\Omega - \int_{\Omega} w \nabla \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega +$$
$$\int_{\Omega} w[(\vartheta + Q)\phi - H - Q_{\phi w}] d\Omega = 0, \quad \forall w \in L_2(\Omega) \tag{8.50}$$

In contrast to the weak form for the divergence form of ADE we restrict the partial integration only to the 2nd-order dispersion term in the convective form of ADE, i.e.,

$$\int_{\Omega} \nabla \cdot [w(\boldsymbol{D} \cdot \nabla\phi)] d\Omega = \int_{\Omega} w \nabla \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla\phi) d\Omega \tag{8.51}$$

By employing the Gauss's integral theorem (2.77) on the LHS term of (8.51) we find

$$\int_\Omega w\nabla\cdot(\boldsymbol{D}\cdot\nabla\phi)d\Omega = \int_\Gamma w(\boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n}\,d\Gamma - \int_\Omega \nabla w\cdot(\boldsymbol{D}\cdot\nabla\phi)d\Omega \tag{8.52}$$

Inserting (8.52) into (8.50) the weak form of the convective form of ADE results

$$\int_\Omega w\acute{\mathcal{R}}\frac{\partial\phi}{\partial t}d\Omega + \int_\Omega w\boldsymbol{q}\cdot\nabla\phi d\Omega + \int_\Omega \nabla w\cdot(\boldsymbol{D}\cdot\nabla\phi)d\Omega +$$
$$\int_\Omega w[(\vartheta + Q)\phi - H - Q_{\phi w}]d\Omega - \int_\Gamma w(\boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n}\,d\Gamma = 0,\quad \forall w\in H^1(\Omega) \tag{8.53}$$

Separating the boundary integral of (8.53) into the three segments $\Gamma = \Gamma_D\cup\Gamma_N\cup\Gamma_C$ imposed by the Dirichlet, Neumann and Cauchy-type BC's, respectively, we invoke the BC's of (8.6) to obtain

$$\int_\Omega w\acute{\mathcal{R}}\frac{\partial\phi}{\partial t}d\Omega + \int_\Omega w\boldsymbol{q}\cdot\nabla\phi d\Omega + \int_\Omega \nabla w\cdot(\boldsymbol{D}\cdot\nabla\phi)d\Omega +$$
$$\int_\Omega w[(\vartheta + Q)\phi - H - Q_{\phi w}]d\Omega - \int_{\Gamma_D} w(\boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n}\,d\Gamma +$$
$$\int_{\Gamma_N} wq_N d\Gamma - \int_{\Gamma_C} w\Phi(\phi_C - \phi)d\Gamma = 0,\quad \forall w\in H^1(\Omega) \tag{8.54}$$

Using this restriction $\forall w\in H_0^1$, the final weak form for the convective form of ADE (8.5) with its BC's (8.6) results

$$\int_\Omega w\acute{\mathcal{R}}\frac{\partial\phi}{\partial t}d\Omega + \int_\Omega w\boldsymbol{q}\cdot\nabla\phi d\Omega + \int_\Omega \nabla w\cdot(\boldsymbol{D}\cdot\nabla\phi)d\Omega +$$
$$\int_\Omega w[(\vartheta + Q)\phi - H]d\Omega + \sum_w w(\boldsymbol{x}_w)\big(\phi_w - \phi(\boldsymbol{x}_w)\big)Q_w(t) +$$
$$\int_{\Gamma_N} wq_N d\Gamma - \int_{\Gamma_C} w\Phi(\phi_C - \phi)d\Gamma = 0,\quad \forall w\in H_0^1(\Omega) \tag{8.55}$$

for solving $\phi \approx \hat{\phi} \in H^1(\Omega)$, where the well-type SPC has been incorporated according to (8.6) (see related discussion in Sect. 8.5.1).

8.5.3 Discussion of Both Weak Forms

We emphasize again that the BC's used in the weak form (8.55) for the convective form of ADE have different meaning in comparison with BC's embodied in the weak form (8.48) for the divergence form of ADE because in general $q_N \neq q_N^\dagger$

and $\Phi \neq \Phi^\dagger$. Only in absence of the normal advective flux $\boldsymbol{q}\cdot\boldsymbol{n} = 0$, it becomes $q_N = q_N^\dagger$ and accordingly $\Phi = \Phi^\dagger$. The consequences on boundary fluxes in both weak forms are obvious. For instance, a natural Neumann BC for the divergence form of ADE $q_N^\dagger = 0$ implies that the boundary segment Γ_N is impervious for the total (both advective and dispersive) flux independent of the actual value of $\boldsymbol{q}\cdot\boldsymbol{n}$, which represents a stronger BC formulation in comparison with the convective form of ADE. On the other hand, a natural Neumann BC for the convective form of ADE $q_N = 0$ ensures at first that the boundary segment Γ_N is only impervious for the dispersive flux, unless $\boldsymbol{q}\cdot\boldsymbol{n} = 0$ can be additionally satisfied. In practical application, the differences between these two weak forms are often not relevant. In solving the convective form of ADE a preceding solution of a flow problem delivers a flow field which satisfies $\boldsymbol{q}\cdot\boldsymbol{n}$ conditions on the boundary in a weak sense and implies appropriate formulations of BC's for both advective and dispersive fluxes in the convective form of ADE, which are equivalent to the divergence form of ADE. In cases, where a total load of a quantity ϕ (consisting of advective plus dispersive fluxes) has to be imposed on a boundary section as formulated by (6.21), (6.28) or (8.7), the Cauchy BC term of (8.55) in the convective form of ADE can be easily utilized as

$$\int_{\Gamma_C} w\Phi(\phi_C - \phi)d\Gamma = -\int_{\Gamma_C} w\boldsymbol{q}\cdot\boldsymbol{n}(\phi_C - \phi)d\Gamma \tag{8.56}$$

where $\boldsymbol{q}\cdot\boldsymbol{n}|_{\Gamma_C}$ is a known advective normal flux on Γ_C so that $\boldsymbol{q}\cdot\boldsymbol{n}\phi_C|_{\Gamma_C}$ prescribes an advective load of quantity ϕ, positive outward-directed on Γ_C.

As discussed in Sect. 6.5.7 outflow BC's (OBC's) can be imposed in two different ways. Commonly, for standard situations a zero-gradient condition, i.e., a natural Neumann BC with $\nabla\phi \approx \mathbf{0}$ is applied. Denoting the boundary portion of the OBC by $\Gamma_{N_O} \subset \Gamma_N \subset \Gamma$, it is specified

$$\int_{\Gamma_{N_O}} wq_N d\Gamma = 0 \quad \text{on} \quad \Gamma_{N_O} \subset \Gamma_N \tag{8.57}$$

for the convective weak form (8.55) and

$$\int_{\Gamma_{N_O}} wq_N^\dagger d\Gamma = \int_{\Gamma_{N_O}} w\phi\boldsymbol{q}\cdot\boldsymbol{n}d\Gamma \quad \text{on} \quad \Gamma_{N_O} \subset \Gamma_N \tag{8.58}$$

for the divergence weak form (8.48). It is obvious, this type of OBC can be simply realized in the convective form, while for the divergence form a surface integral remains to be treated implicitly because ϕ is unknown and the normal flux $\boldsymbol{q}\cdot\boldsymbol{n}$ must be determined (or be known) on Γ_{N_O}. The second and alternative way is to impose the OBC fully implicitly, even for the gradient-driven dispersive boundary flux. Using this BC formulation the Neumann-type boundary integrals are replenished to specify the *implicit* OBC

$$\int_{\Gamma_{N_O}} wq_N \, d\Gamma = -\int_{\Gamma_{N_O}} w(\boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n} \, d\Gamma \quad \text{on} \quad \Gamma_{N_O} \subset \Gamma_N \tag{8.59}$$

for the convective weak form (8.55) and

$$\int_{\Gamma_{N_O}} wq_N^\dagger \, d\Gamma = \int_{\Gamma_{N_O}} w(\phi\boldsymbol{q} - \boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n} \, d\Gamma \quad \text{on} \quad \Gamma_{N_O} \subset \Gamma_N \tag{8.60}$$

for the divergence weak form (8.48), which must be treated with unknown ϕ. We conclude that OBC requires in general an implicit treatment of the specific Neumann-type surface integrals for the divergence weak form, which is more complex. In contrast, however, the OBC in the convective weak form can be simply specified, unless the zero-gradient Neumann condition on the outflow boundary is not appropriate under specific situations (cf. discussion in Sect. 6.5.7).

8.6 Spatial Discretization by Finite Elements

The governing weak forms derived in Sect. 8.5 contains integral expressions which have to be solved. To accomplish an approximate solution via FEM the continuum domain with its boundary $\bar{\Omega} = \Omega \cup \Gamma$ is subdivided into a set of nonoverlapping subdomains, called *finite elements* (see Fig. 8.5), such that

$$\bar{\Omega} \approx \hat{\bar{\Omega}} \equiv \bigcup_{e=1}^{N_\mathrm{E}} \bar{\Omega}^e \quad \text{with} \quad \bar{\Omega}^e = \Omega^e \cup \Gamma^e, \quad \bar{\Omega}^e \neq \emptyset \tag{8.61}$$

where N_E is the number of finite elements, $\hat{\bar{\Omega}}$ is the approximate global domain, Ω^e and Γ^e are the domain and the boundary of each finite element e, respectively. The *basic idea* of the finite element spatial discretization (8.61) is to split any integral that appears in the weak statements into a sum over the elements

$$\begin{aligned} \textstyle\int_\Omega \{\ldots\} d\Omega &= \sum_{e=1}^{N_\mathrm{E}} \int_{\Omega^e} \{\ldots\} d\Omega^e \\ \textstyle\int_\Gamma \{\ldots\} d\Gamma &= \sum_{e=1}^{N_\mathrm{E}} \int_{\Gamma^e} \{\ldots\} d\Gamma^e \end{aligned} \tag{8.62}$$

This nonoverlapping sum over all elements is called *assembly*. The actual domain $\hat{\bar{\Omega}}$ assembled by all these elements $\bigcup_e (\Omega^e \cup \Gamma^e)$ is termed *finite element mesh* (Fig. 8.5). The goal of the assembly (8.62) is to accomplish an easily tractable, sufficiently accurate and efficient integration on element level. This can be attained

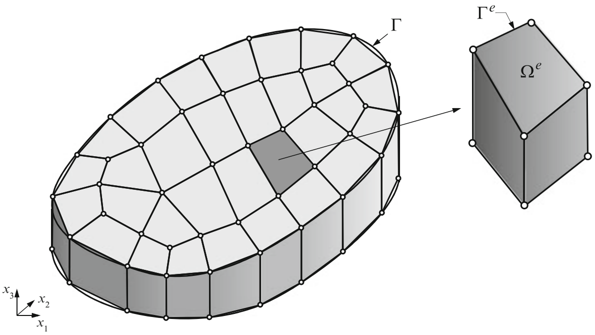


Fig. 8.5 Spatial discretization of a continuum domain with its boundary $\bar{\Omega} = \Omega \cup \Gamma$ by finite elements $\bar{\Omega}^e = \Omega^e \cup \Gamma^e$, $(e = 1, \dots, N_{\mathrm{E}})$ forming a finite element mesh

by choosing suitable shapes for the finite element $\bar{\Omega}^e$ that have appropriate geometric entities (vertices, mid-sides) to match the interpolation for the approximate solution $\hat{\phi}$ according to (8.16) with a desired accuracy. The finite element $\bar{\Omega}^e$ can be a line, triangle or quadrilateral in 1D, 2D or 3D, respectively, and the degree of interpolation over it can be linear, quadratic or even higher. In practice, the phrase *finite element* refers to both the geometry of the element and degree of approximation used for the solution variable(s), e.g., a quadratic quadrilateral element is a 2D quadrilateral shape with a biquadratic (biparabolic) interpolation, a linear triangular prismatic element represents a 3D pentahedral shape with trilinear interpolation, and so forth. Commonly used finite elements in 1D, 2D and 3D are depicted in Fig. 8.6.

8.7 Elementwise Continuous Approximations

The assembly (8.62) of the finite elements is only valid if the basis (interpolation) functions (8.16) satisfy requirements on continuity. The basis functions have to be restricted to avoid any infinite terms in the integrals of the approximate weak statement. The situation is explained in Fig. 8.7. Let us consider the interfacing boundary of two adjacent finite elements, where we study the approximate function $\hat{\phi}$ and its derivatives in a very small distance $\delta \rightarrow 0$. Within the elementwise interpolation procedure we can ensure that $\hat{\phi}$ is continuous everywhere in $\hat{\bar{\Omega}}$ and also at the element interface(s). However, this must not be the case for the first derivative anymore, which can become discontinuous at element interfaces. While the first derivative is discontinuous, its value remains in a finite value and any integrand of the weak form containing up to a first-order derivative is finite and accordingly evaluable. In contrast, however, consider its second derivative, which tends to an infinite value at the element interface. Such a term is no more square integrable and the assembly (8.62) fails.

The continuity requirement can be generalized as follows. Suppose the integrand in the approximate weak statement contains up to $(m + 1)$th derivatives, then

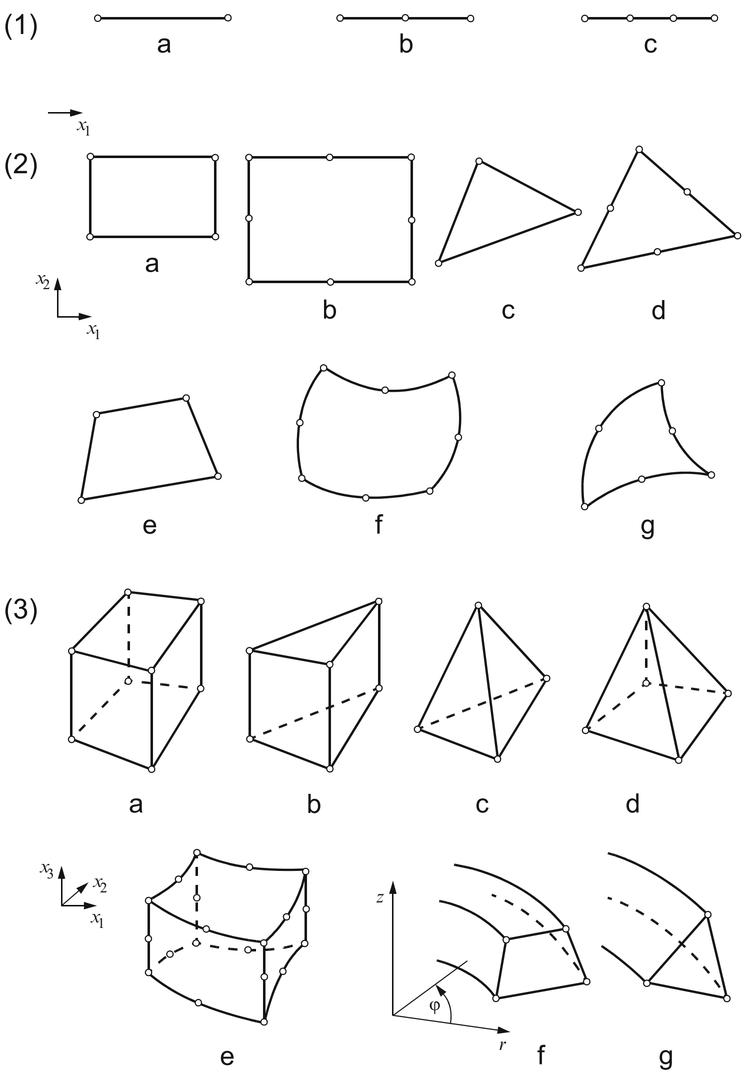


Fig. 8.6 Overview of commonly used finite elements. (**1**) 1D elements: (*a*) linear, (*b*) quadratic, (*c*) cubic; (**2**) 2D elements: (*a*) linear rectangular, (*b*) quadratic rectangular, (*c*) linear triangular, (*d*) quadratic triangular, (*e*) linear quadrilateral, (*f*) quadratic curved quadrilateral, (*g*) quadratic curved triangular; (**3**) 3D elements: (*a*) linear quadrilateral prism (hexahedron), (*b*) linear triangular prism (pentahedron), (*c*) linear tetrahedron, (*d*) linear pyramid, (*e*) quadratic curved hexahedron, (*f*) axisymmetric linear rectangular ring, (*g*) axisymmetric linear triangular ring (Modified from [76])

continuity in the mth derivative of the approximate function must be satisfied. This is called the C_m-continuity requirement. The validity of the assembly (8.62) requires the fulfillment of the C_m-continuity in any finite element basis function.

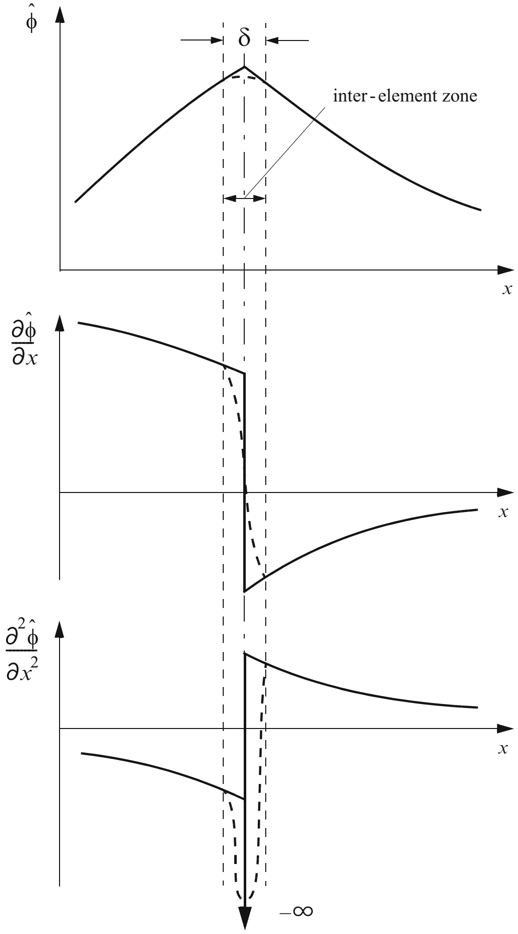

$\hat{\phi}$
δ
inter - element zone
x
$\frac{\partial\hat{\phi}}{\partial x}$
x
$\frac{\partial^2\hat{\phi}}{\partial x^2}$
x
$-\infty$

Fig. 8.7 Inter-element behavior of C_0 continuous approximate function $\hat{\phi}$ and its derivatives: While $\hat{\phi}$ is continuous at the element interface, its first derivative becomes discontinuous within the inter-element zone $\delta \to 0$, but is still finite. The second derivative, however, may become infinite (Modified from [590])

Now, having a look to the weak forms as derived in Sect. 8.5, we recognize that the highest derivatives are only of first order (thanks to the reduction of the 2nd-order derivatives in the dispersion term due to applying the Gauss's integral theorem). Hence, it is sufficient to satisfy only C_0—continuity in the interpolation function(s) of the unknown variable(s), i.e., the element basis functions $\hat{\phi}$ have to be chosen in such a way that the zero derivatives are continuous and their first derivatives, while discontinuous at the element interfaces (they actually suffer jumps at nodal points), need only to be square integrable.

The most important class of C_0 basis functions refers to *Lagrangian polynomials*, which are standard in FEM. C_0 functions are commonly sufficient for all problems of advection-dispersion type, which are encountered in the present flow and transport processes. On the other hand, a higher order continuity, e.g., C_1 functions satisfying continuity of both zero and first derivatives, can be provided by Hermitian polynomials [173, 280]. Although C_1 Hermitian polynomials can achieve a higher accuracy for the first derivatives, however, at the expense of additional degrees of freedom associated with computational extra costs, their practical applicability has shown limited (e.g., to undistorted elements) and rather cumbersome. Indeed, we need not a continuity higher than C_0. In the following we exclusively prefer C_0 continuous basis functions for various element types in 1D, 2D and 3D.

8.8 Finite Element Basis Functions

8.8.1 Shape Function, Master Element and Isoparametric Element Type

In using assembly (8.62) it is advantageous to restrict the interpolation of the unknown variable(s) within each finite element $\bar{\Omega}^e = \Omega^e \cup \Gamma^e$, such that the approximation $\hat{\phi}(\boldsymbol{x}, t)$ according to (8.16) can then be formed as the *union* of the finite element approximations $\hat{\phi}^e(\boldsymbol{x}^e, t)$ on $\bar{\Omega}^e$, viz.,

$$\phi(\boldsymbol{x}, t) \approx \hat{\phi}(\boldsymbol{x}, t) = \bigcup_{e=1}^{N_{\mathrm{E}}} \hat{\phi}^e(\boldsymbol{x}^e, t) \tag{8.63}$$

Note that it is not possible to simply sum $\hat{\phi}^e$ over e since a double contribution would occur on every finite element boundary. Thus, a summation without overlap of element boundary will be indicated by the union symbol:

$$\bigcup_{e=1}^{N_{\mathrm{E}}} (\ldots) = \sum_{e=1}^{N_{\mathrm{E}}} (\ldots) \quad \text{without boundary overlap} \tag{8.64}$$

On any finite element domain $\bar{\Omega}^e$, the generic form for $\hat{\phi}^e$ is

$$\hat{\phi}^e(\boldsymbol{x}^e, t) = \sum_{J=1}^{N_{\mathrm{BN}}} N_J^e(\boldsymbol{x}^e)\, \phi_J^e(t) \tag{8.65}$$

where ϕ_J^e are the set of unknown coefficients at the nodes J belonging to the element e and $N_J^e(\boldsymbol{x}^e)$ are the set of given C_0 continuous basis functions, called *shape functions*, associated with the element e and the *local* node number J (note that we shall differ between local and global node numbering as further discussed below). The element shape functions $N_J^e(\boldsymbol{x}^e)$ represent polynomials of 1st, 2nd or even higher degree. In practice, however, we prefer polynomials of 1st degree and, optionally, 2nd degree. There are as many of these polynomials as there are nodal points N_{BN} in $\bar{\Omega}^e$. To achieve a continuous representation of $\hat{\phi}$ (cf. Sect. 8.7) the element shape functions must satisfy C_0-continuity, for which the approximate solution is continuous and have piecewise continuous first-order derivatives. Those element shape functions are referred to as C_0-class elements, which will be generally used in the following.

The element shape functions have the following property at the nodal points:

$$N_J^e(\boldsymbol{x}_I^e) = \begin{cases} \delta_{IJ} & \text{for} \quad \boldsymbol{x}_I^e \in \bar{\Omega}^e \\ 0 & \text{otherwise} \end{cases} \tag{8.66}$$

where δ_{IJ} is the Kronecker symbol (2.7) and $\boldsymbol{x}_I^e$ are the Cartesian coordinates of local node I (cf. (2.30)). From (8.66) it directly follows that

$$\sum_{J=1}^{N_{\text{BN}}} N_J^e(\boldsymbol{x}^e) = 1, \quad \forall \boldsymbol{x}^e \in \bar{\Omega}^e \tag{8.67}$$

The ability of handling nonuniform and distorted geometries is an important feature of the FEM. A fundamental aspect of FEM is the use of a *master element* $\bar{\Omega}_m^e = \Omega_m^e \cup \Gamma_m^e$, where all element-related inner products and integrations are performed in local coordinates $\boldsymbol{\eta}$ defined as

$$\boldsymbol{\eta}^T = \begin{cases} (\xi \; \eta \; \zeta) & \text{3D} \\ (\xi \; \eta) & \text{2D and axisymmetric} \\ (\xi) & \text{1D} \end{cases} \tag{8.68}$$

A one-to-one mapping (coordinate transformation, see Sect. 2.1.5) bridges the global Euclidean $\boldsymbol{x}-$space and the local (computational) $\boldsymbol{\eta}-$space of the master element $\bar{\Omega}_m^e$:

$$\boldsymbol{x}^e = \boldsymbol{x}^e(\boldsymbol{\eta}) \tag{8.69}$$

The element geometry of the master element $\bar{\Omega}_m^e$ is always Cartesian (rectangular) so that the integration on such an element level can be efficiently computed. Based on this mapping the finite elements can be distorted easily to fit most applicable geometries (Fig. 8.8). For this purpose it is advantageous to define the element shape functions in their local coordinates $N_J^e(\boldsymbol{\eta})$, such that (8.65) becomes

$$\hat{\phi}^e(\boldsymbol{x}^e(\boldsymbol{\eta}), t) = \sum_{J=1}^{N_{\text{BN}}} N_J^e(\boldsymbol{\eta})\, \phi_J^e(t) \tag{8.70}$$

and global coordinates $\boldsymbol{x}$ are related to the local coordinates $\boldsymbol{\eta}$ by using the interpolation

$$\boldsymbol{x}^e = \sum_{J=1}^{N_{\text{X}}} N_J^e(\boldsymbol{\eta})\, \boldsymbol{x}_J^e \tag{8.71}$$

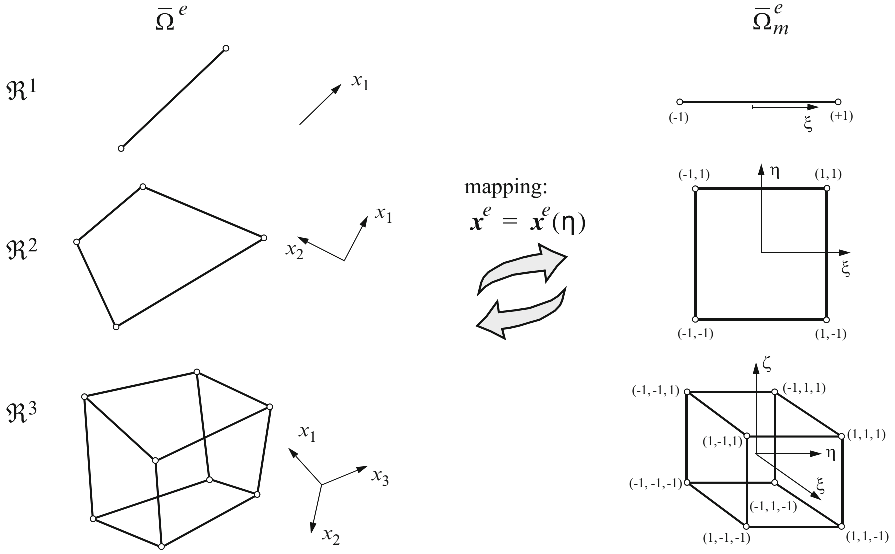


Fig. 8.8 Finite elements with one-to-one mapping onto $\Re^D$ $(D = 1, 2, 3)$

with

$$\boldsymbol{x} = \bigcup_{e=1}^{N_E} \boldsymbol{x}^e \tag{8.72}$$

where N_X is the number of element polynomials used for the geometry interpolation and $\boldsymbol{x}_j^e$ are the global coordinates of node j on element e. According to the choice of N_X it is distinguished into (1) *isoparametric elements* with $N_X = N_{BN}$, i.e., polynomial approximation is used for both geometry and variables, (2) superparametric elements with $N_X > N_{BN}$, where a higher order approximation is used for the geometry, and (3) subparametric elements with $N_X < N_{BN}$, where a lower order approximation is used for the geometry compared to the variable approximation. Most efficient and ideal for our needs are isoparametric elements, which will be generally preferred in the present FEM. Appendix G summarizes the isoparametric finite elements used in FEFLOW for 1D, 2D (incl. axisymmetric) and 3D problems.

8.8.2 Local and Global Shape Functions

To illustrate the construction of finite element basis functions let us consider at first the simplest case: the use of linear isoparametric shape functions in a 1D geometry $x \in \Re^1$ (see also Table G.1 in Appendix G). Figure 8.9 displays the master element $\bar{\Omega}_m^e$ with the local node numbering $J = 1, 2$, the linear shape functions expressed

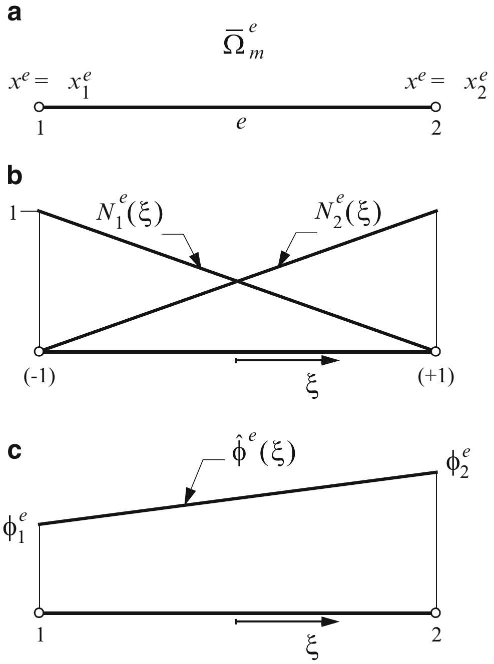


Fig. 8.9 Piecewise-linear shape functions for 1D element: (**a**) master element $\bar{\Omega}_m^e$ with local node numbering, (**b**) shape functions $N_J^e(\xi)$ $(J = 1, 2)$ in local coordinate $-1 \leq \xi \leq 1$, (**c**) approximate variable $\hat{\phi}^e(\xi)$ as linear function over element e

in the local coordinate $(-1 \leq \xi \leq 1)$

$$N_1^e(\xi) = \tfrac{1}{2}(1 - \xi) \qquad N_2^e(\xi) = \tfrac{1}{2}(1 + \xi) \tag{8.73}$$

and the resulting approximate function $\hat{\phi}^e$ over the element e. Using the mapping relation (8.71) for the linear 2-node element

$$x^e = \sum_{J=1}^{N_{\mathrm{BN}}=2} N_J^e(\xi)\, x_J^e \tag{8.74}$$

we find $\xi = (2x^e - x_1^e - x_2^e)/(x_2^e - x_1^e)$ and the shape functions can also be written in the global coordinate x^e, viz.,

$$N_1^e(x^e) = \frac{x_2^e - x^e}{x_2^e - x_1^e} \qquad N_2^e(x^e) = \frac{x^e - x_1^e}{x_2^e - x_1^e} \tag{8.75}$$

where x_1^e and x_2^e are the $x-$coordinates of local node number 1 and 2, respectively, of element e. Then, the approximate variable $\hat{\phi}^e$ is linear over element e (Fig. 8.9c):

$$\begin{aligned} \hat{\phi}^e(\xi) &= \tfrac{1}{2}\big[(\phi_2^e - \phi_1^e)\xi + \phi_1^e + \phi_2^e\big] \qquad \text{or} \\ \hat{\phi}^e(x^e) &= \frac{1}{x_2^e - x_1^e}\big[(\phi_2^e - \phi_1^e)x^e + x_2^e\phi_1^e - x_1^e\phi_2^e\big] \end{aligned} \tag{8.76}$$

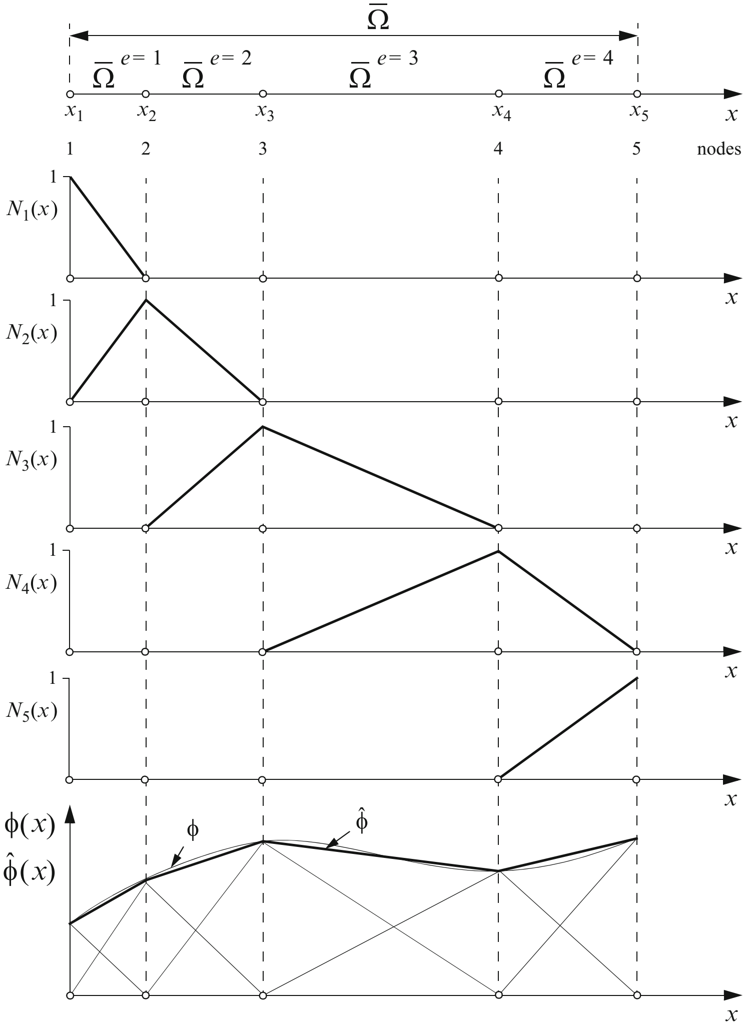


Fig. 8.10 Example of 1D finite element mesh consisting of four linear elements. Display of global basis function $N_j(x)$, $(j = 1, \ldots, 5)$ and linear approximation of variable ϕ by $\hat{\phi}$

Now, let us consider a 1D mesh consisting of four linear elements and five global nodes as shown in Fig. 8.10. The approximate function $\hat{\phi}$ is represented using global shape functions N_j that are equal to one at node j and zero at all other nodes. Accordingly, the global function $\hat{\phi}$ in the mesh shown in Fig. 8.10 can be written as

$$\hat{\phi}(x,t) = \sum_{j=1}^{N_{\mathrm{P}}=5} N_j(x)\phi_j(t) = \bigcup_{e=1}^{N_{\mathrm{E}}=4} \hat{\phi}^e(x,t) \tag{8.77}$$

with

$$\hat{\phi}^e(x,t) = \sum_{J=1}^{N_{\mathrm{BN}}=2} N_J^e(\xi)\,\phi_J^e \tag{8.78}$$

It can be observed that, as j and J vary, all elements contain shape functions, which are similar in global coordinates x and identical in the local coordinates ξ. This is a key issue for devising efficient operations on the master element level, which are generic and will be performed widely independent of the global (physical) coordinates as discussed further below.

Now, we have to emphasize the difference between the global node numbering used in N_j and the local node numbering used in N_J^e and to consider how we can relate properties between the global and the local systems: *Any uppercase nodal index J associated with an element-rank quantity represents a local node number, while a lowercase nodal index j of a quantity without element rank means a global node number*. For the mesh of Fig. 8.10 it is seen that

$$\begin{array}{llll} N_1^1(x^1) = N_1(x), & N_2^1(x^1) = N_2(x) & \text{over element} & e = 1 \\ N_1^2(x^2) = N_2(x), & N_2^2(x^2) = N_3(x) & \text{over element} & e = 2 \\ N_1^3(x^3) = N_3(x), & N_2^3(x^3) = N_4(x) & \text{over element} & e = 3 \\ N_1^4(x^4) = N_4(x), & N_2^4(x^4) = N_5(x) & \text{over element} & e = 4 \end{array} \tag{8.79}$$

which can be written in a matrix form as follows

$$\begin{pmatrix} N_1 \\ N_2 \end{pmatrix}^e = \boldsymbol{\Delta}^e \cdot \begin{pmatrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \end{pmatrix}, \qquad e = 1, 2, \ldots, N_{\mathrm{E}} \tag{8.80}$$

or more generally

$$N_J^e = \sum_{j=1}^{N_{\mathrm{P}}} \Delta_{Jj}^e N_j, \qquad J = 1, 2, \ldots, N_{\mathrm{BN}} \tag{8.81}$$

where $\boldsymbol{\Delta}^e = \Delta_{Jj}^e$ is the *Boolean matrix* of element e having the property:

$$\Delta_{Jj}^e = \begin{cases} 1 & \text{if the local node } J \text{ corresponds to the global node } j \\ 0 & \text{otherwise} \end{cases} \tag{8.82}$$

The Boolean matrix $\boldsymbol{\Delta}^e$ will prove to be convenient in derivations of finite element equations, where local quantities have to be related to properties of the global coordinate system. The Boolean matrices for the present mesh of Fig. 8.10 result for example:

$$\boldsymbol{\Delta}^1 = \begin{pmatrix} 1\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0 \end{pmatrix}, \ \boldsymbol{\Delta}^2 = \begin{pmatrix} 0\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0 \end{pmatrix}, \ \boldsymbol{\Delta}^3 = \begin{pmatrix} 0\,0\,1\,0\,0 \\ 0\,0\,0\,1\,0 \end{pmatrix}, \ \boldsymbol{\Delta}^4 = \begin{pmatrix} 0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,1 \end{pmatrix} \tag{8.83}$$

Using the Boolean matrix (8.82), the global shape functions $N_j(\boldsymbol{x})$ appearing in the global approximate solution

$$\hat{\phi}(\boldsymbol{x}, t) = \sum_{j=1}^{N_\mathrm{P}} N_j(\boldsymbol{x})\phi_j(t) \tag{8.84}$$

can be directly expressed by local shape functions N_J^e according to

$$N_j(\boldsymbol{x}) = \bigcup_{e=1}^{N_\mathrm{E}} \left(\sum_{J=1}^{N_\mathrm{BN}} N_J^e(\boldsymbol{\eta})\, \Delta_{Jj}^e \right) \tag{8.85}$$

To increase the accuracy of interpolation a quadratic shape function rather than a linear shape function can be chosen. Quadratic interpolation functions are generated by adding an additional node at the midside of each element as shown in Fig. 8.11 for 1D geometry. The shape functions for this quadratic 3-node element are (cf. also Table G.1 in Appendix G)

$$N_1^e(\xi) = \tfrac{1}{2}\xi(\xi - 1) \qquad N_2^e(\xi) = 1 - \xi^2 \qquad N_3^e(\xi) = \tfrac{1}{2}\xi(\xi + 1) \tag{8.86}$$

Apart from the different polynomials and the number of polynomials (= number of nodes N_BN) per master element appearing for the quadratic element type, the construction of the basis function is based on the same principles as stated above for the linear element. The same is also true for isoparametric elements in higher dimensions (see Tabs. G.2–G.4 in Appendix G for the family of 2D and 3D elements used in FEFLOW). An example of a 2D triangle mesh for a piecewise bilinear approximation of $\hat{\phi}$ is shown in Fig. 8.12. The shape functions of each triangle for the three nodes written in the local coordinates ($0 \leq \xi, \eta \leq 1$) are

$$N_1^e(\xi, \eta) = 1 - \xi - \eta \qquad N_2^e(\xi, \eta) = \xi \qquad N_3^e(\xi, \eta) = \eta \tag{8.87}$$

Furthermore, using the mapping relation (8.71) as

$$x^e = \sum_{J=1}^{N_\mathrm{BN}=3} N_J^e(\xi, \eta)\, x_J^e \qquad y^e = \sum_{J=1}^{N_\mathrm{BN}=3} N_J^e(\xi, \eta)\, y_J^e \tag{8.88}$$

written with (8.87) as

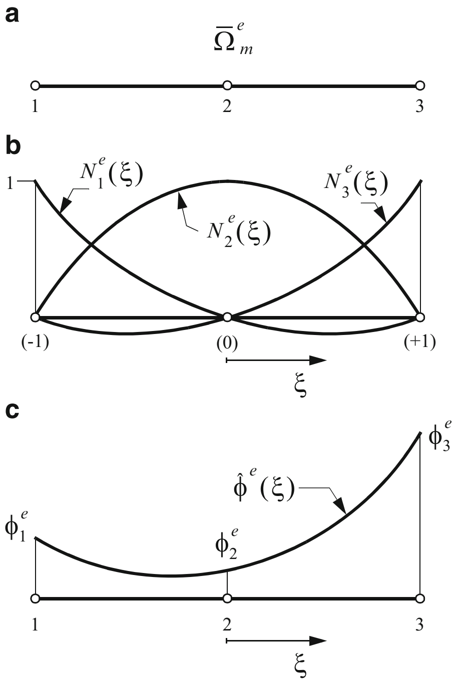


Fig. 8.11 Piecewise-quadratic shape functions for 1D element: (**a**) master element $\bar{\Omega}^e_m$ with local node numbering, (**b**) shape functions $N^e_J(\xi)$ $(J = 1, 2, 3)$ in local coordinate $-1 \leq \xi \leq 1$, (**c**) approximate variable $\hat{\phi}^e(\xi)$ as quadratic function over element e

$$\begin{aligned} \begin{pmatrix} x \\ y \end{pmatrix}^e &= \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}^e + \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}^e \cdot \begin{pmatrix} \xi \\ \eta \end{pmatrix} \\ &= \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}^e + \underbrace{\begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}^e}_{\boldsymbol{J}^e} \cdot \begin{pmatrix} \xi \\ \eta \end{pmatrix} \end{aligned} \tag{8.89}$$

we can express the local coordinates as

$$\begin{aligned} \begin{pmatrix} \xi \\ \eta \end{pmatrix} &= \frac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} J_{22} & -J_{12} \\ -J_{21} & J_{11} \end{pmatrix}^e \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}^e \\ &= \frac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix}^e \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}^e \end{aligned} \tag{8.90}$$

and find finally the shape functions of the linear triangle in global coordinates according to

$$\begin{aligned} N^e_1(x^e, y^e) &= \frac{1}{|\boldsymbol{J}^e|} \big[x_2 y_3 - x_3 y_2 + (y_2 - y_3)x + (x_3 - x_2)y \big]^e \\ N^e_2(x^e, y^e) &= \frac{1}{|\boldsymbol{J}^e|} \big[x_3 y_1 - x_1 y_3 + (y_3 - y_1)x + (x_1 - x_3)y \big]^e \\ N^e_3(x^e, y^e) &= \frac{1}{|\boldsymbol{J}^e|} \big[x_1 y_2 - x_2 y_1 + (y_1 - y_2)x + (x_2 - x_1)y \big]^e \end{aligned} \tag{8.91}$$

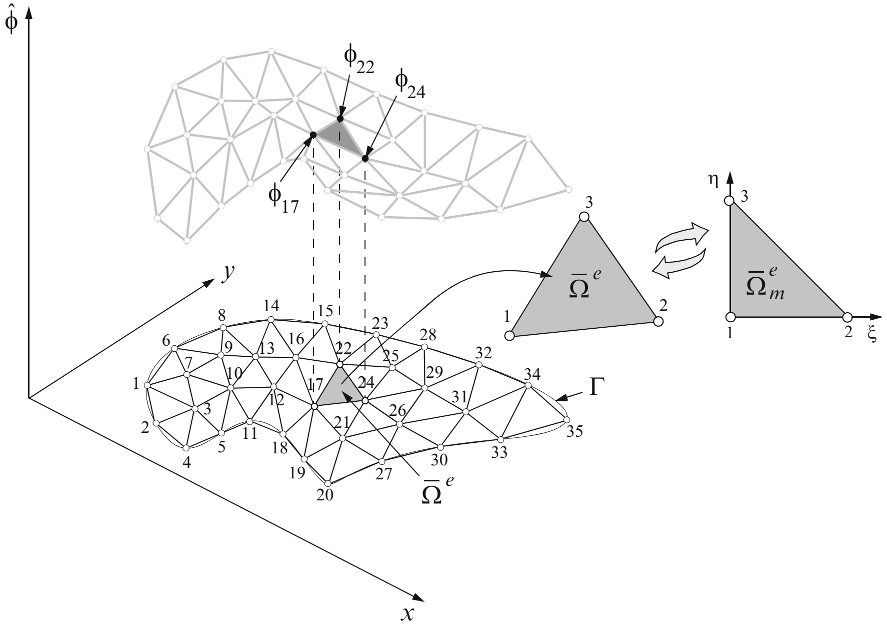


Fig. 8.12 2D triangular element mesh with global node numbering and a piecewise linear approximation of $\hat{\phi}$. Selected triangular element $\bar{\Omega}^e \subset \Re^2$ with local node numbering and mapping onto master element $\bar{\Omega}^e_m$ in local coordinates ξ and η

where the determinant of the Jacobian $\boldsymbol{J}^e$ (equal to twice the area of triangle) is given by

$$|\boldsymbol{J}^e| = [J_{11}J_{22} - J_{21}J_{12}]^e = [x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(y_1 - y_2)]^e \tag{8.92}$$

Then, the approximate variable $\hat{\phi}$ forms a piecewise-bilinear function over the solution domain as exemplified in Fig. 8.12

$$\hat{\phi}(x, y, t) = \sum_j N_j(x, y)\,\phi_j(t) \tag{8.93}$$

with $N_j(x, y) = \bigcup_e (\sum_J N^e_J(\xi, \eta)\,\Delta^e_{Jj})$.

In the same way we are able to construct finite element basis functions for all element types we have in mind for 1D, 2D and 3D applications. The family of finite elements preferred in FEFLOW are summarized in Appendix G. While the finite element basis functions $N^e_J(\boldsymbol{\eta})$ are expressed in analytical forms, the required coordinate transformation (mapping) between $\bar{\Omega}^e$ in the global coordinate system and $\bar{\Omega}^e_m$ in the local coordinate system represents a generic task. It can be performed by very efficient basis operations for each element, which will be thoroughly described next.

8.9 Galerkin Finite Element Weak Statement

Using the weak forms derived for the divergence and convective form of ADE according to (8.48) and (8.55), respectively, their weak statements (8.36) applied to the approximate variable $\hat{\phi}$ result

$$\begin{aligned}\text{WS} = \int_{\Omega} w_i \frac{\partial(\mathcal{R}\hat{\phi})}{\partial t} d\Omega - \int_{\Omega} \hat{\phi}\boldsymbol{q} \cdot \nabla w_i d\Omega + \int_{\Omega} \nabla w_i \cdot (\boldsymbol{D} \cdot \nabla\hat{\phi}) d\Omega + \\ \int_{\Omega} w_i (\vartheta\hat{\phi} - H) d\Omega + \phi_w Q_w(t)\big|_i + \int_{\Gamma_N} w_i q_N^{\dagger} d\Gamma - \\ \int_{\Gamma_C} w_i \Phi^{\dagger} (\phi_C - \hat{\phi}) d\Gamma = 0 \quad \forall w_i \in H_0^1(\Omega), \;\; 1 \leq i \leq N_{\text{EQ}}\end{aligned} \tag{8.94}$$

for the divergence form of ADE and

$$\begin{aligned}\text{WS} = \int_{\Omega} w_i \acute{\mathcal{R}} \frac{\partial\hat{\phi}}{\partial t} d\Omega + \int_{\Omega} w_i \boldsymbol{q} \cdot \nabla\hat{\phi} d\Omega + \int_{\Omega} \nabla w_i \cdot (\boldsymbol{D} \cdot \nabla\hat{\phi}) d\Omega + \\ \int_{\Omega} w_i [(\vartheta + Q)\hat{\phi} - H] d\Omega + \big(\phi_w - \phi(\boldsymbol{x}_w)\big) Q_w(t)\big|_i + \int_{\Gamma_N} w_i q_N d\Gamma - \\ \int_{\Gamma_C} w_i \Phi (\phi_C - \hat{\phi}) d\Gamma = 0 \quad \forall w_i \in H_0^1(\Omega), \;\; 1 \leq i \leq N_{\text{EQ}}\end{aligned} \tag{8.95}$$

for the convective form of ADE.

Now, we discretize the domain Ω and its boundary Γ by finite elements via (8.62), introduce the semidiscrete finite element basis function for $\hat{\phi} = \hat{\phi}(\boldsymbol{x}, t)$ over each element e

$$\begin{aligned}\hat{\phi}(\boldsymbol{x}, t) = \sum_{j=1}^{N_{\text{P}}} N_j(\boldsymbol{x}) \phi_j(t) = \bigcup_{e=1}^{N_{\text{E}}} \hat{\phi}^e(\boldsymbol{x}^e, t) \\ \hat{\phi}^e(\boldsymbol{x}^e, t) = \sum_{j=1}^{N_{\text{P}}} \sum_{J=1}^{N_{\text{BN}}} N_J^e(\boldsymbol{\eta})\, \Delta_{Jj}^e \phi_j(t)\end{aligned} \tag{8.96}$$

and choose the Galerkin method (Table 8.1), where the weighting function becomes identical to the basis function[5]

[5] The weak statements (8.94) and (8.95) imply that the weighting functions w_i belong to the H_0^1 functional space (8.23). On the other hand, the basis functions N_i belong to the H^1 functional space, i.e., they do not vanish on Dirichlet boundaries: $N_i \neq 0$ on Γ_D. Nevertheless, we may use WS in form of (8.94) and (8.95) with $w_i = N_i \in H^1(\Omega), \;\; 1 \leq i \leq N_{\text{P}}$, where we enforce at first a zero flux $(\phi\boldsymbol{q} - \boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n} \approx 0$ or $-(\boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n} \approx 0$ on Γ_D in the original weak statements (8.47) and (8.54), respectively, and incorporate the actual Dirichlet (essential) BC's afterwards via a direct manipulation of the resulting discrete matrix system as further discussed in Sect. 8.16.

$$w_i(\boldsymbol{x}) = N_i(\boldsymbol{x}) = \bigcup_{e=1}^{N_\mathrm{E}} \Bigg(\sum_{I=1}^{N_\mathrm{BN}} N_I^e(\boldsymbol{\eta})\, \Delta_{Ii}^e \Bigg) \tag{8.97}$$

we find the following finite element forms of the Galerkin weak statement (GWS)

$$\begin{aligned}
\mathrm{GWS} = {} & \sum_e \int_{\Omega^e} N_i \frac{\partial}{\partial t} \big[\mathcal{R}(\sum_j N_j \phi_j) \big] d\Omega^e - \sum_e \int_{\Omega^e} (\sum_j N_j \phi_j) \boldsymbol{q} \cdot \nabla N_i \, d\Omega^e + \\
& \sum_e \int_{\Omega^e} \nabla N_i \cdot \big[\boldsymbol{D} \cdot \nabla (\sum_j N_j \phi_j) \big] d\Omega^e + \sum_e \int_{\Omega^e} N_i \big[\vartheta (\sum_j N_j \phi_j) - H \big] d\Omega^e + \\
& \phi_w Q_w(t)\big|_i + \sum_e \int_{\Gamma_N^e} N_i q_N^\dagger \, d\Gamma^e - \sum_e \int_{\Gamma_C^e} N_i \varPhi^\dagger \big[\phi_C - (\sum_j N_j \phi_j) \big] d\Gamma^e = 0 \\
& \qquad 1 \le i, j \le N_\mathrm{P}
\end{aligned} \tag{8.98}$$

for the divergence form of ADE and

$$\begin{aligned}
\mathrm{GWS} = {} & \sum_e \int_{\Omega^e} N_i \acute{\mathcal{R}} \frac{\partial}{\partial t} (\sum_j N_j \phi_j) d\Omega^e + \sum_e \int_{\Omega^e} N_i \boldsymbol{q} \cdot \nabla (\sum_j N_j \phi_j) d\Omega^e + \\
& \sum_e \int_{\Omega^e} \nabla N_i \cdot \big[\boldsymbol{D} \cdot \nabla (\sum_j N_j \phi_j) \big] d\Omega^e + \sum_e \int_{\Omega^e} N_i \big[(\vartheta + Q)(\sum_j N_j \phi_j) - H \big] d\Omega^e + \\
& (\phi_w - \phi_i) Q_w(t)\big|_i + \sum_e \int_{\Gamma_N^e} N_i q_N \, d\Gamma^e - \\
& \sum_e \int_{\Gamma_C^e} N_i \varPhi \big[\phi_C - (\sum_j N_j \phi_j) \big] d\Gamma^e = 0 \quad 1 \le i, j \le N_\mathrm{P}
\end{aligned} \tag{8.99}$$

for the convective form of ADE. The indicated integrals in (8.98) and (8.99) are evaluated at the element level e and *assembled* (summed up) into a global matrix system of the form

$$\boldsymbol{O} \cdot \dot{\boldsymbol{\phi}} + \boldsymbol{K} \cdot \boldsymbol{\phi} - \boldsymbol{F} = \boldsymbol{0} \tag{8.100}$$

The assembly process in forming the global matrices and vectors from element contributions will be described more in detail in Sect. 8.10. In (8.100) $\boldsymbol{\phi} = \boldsymbol{\phi}(t)$ is a column vector of the state-variable approximation coefficients

$$\boldsymbol{\phi} = \phi_j = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N_\mathrm{P}} \end{pmatrix} \tag{8.101}$$

to be solved as unknowns from the resulting equation system (8.100). The superposed dot in (8.100) means differentiation with respect to time t

$$\dot{\phi} = \frac{d}{dt}\phi(t) = \frac{d}{dt}\phi_j(t) = \begin{pmatrix} \frac{d\phi_1}{dt} \\ \frac{d\phi_2}{dt} \\ \vdots \\ \frac{d\phi_{N_\mathrm{P}}}{dt} \end{pmatrix} \tag{8.102}$$

The components of the global rank square matrices $\boldsymbol{O}$ and $\boldsymbol{K}$ as well as the global column vector $\boldsymbol{F}$ are written in indicial notation as[6]

[6] We can alternatively write the matrices and vectors in using directly the global shape function (8.85) with the global node numbers i, j:

$$\begin{aligned}
O_{ij} &= \begin{cases} \sum_e \int_{\Omega^e} \mathcal{R}^e \, N_i N_j \, d\Omega^e & \text{divergence form} \\ \sum_e \int_{\Omega^e} \acute{\mathcal{R}}^e \, N_i N_j \, d\Omega^e & \text{convective form} \end{cases} \\
A_{ij} &= \begin{cases} -\sum_e \int_{\Omega^e} \boldsymbol{q}^e \cdot \nabla N_i N_j \, d\Omega^e & \text{divergence form[7]} \\ \sum_e \int_{\Omega^e} N_i \boldsymbol{q}^e \cdot \nabla N_j \, d\Omega^e & \text{convective form} \end{cases} \\
C_{ij} &= \sum_e \int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{D}^e \cdot \nabla N_j) d\Omega^e \\
R_{ij} &= \begin{cases} \sum_e \int_{\Omega^e} (\vartheta^e + \frac{\partial \mathcal{R}^e}{\partial t}) N_i N_j \, d\Omega^e & \text{divergence form} \\ \sum_e \int_{\Omega^e} (\vartheta^e + Q^e) N_i N_j \, d\Omega^e - \delta_{ij} Q_w(t)|_i & \text{convective form} \end{cases} \\
B_{ij} &= \begin{cases} \sum_e \Big(\int_{\Gamma_C^e} \Phi^{\dagger^e} N_i N_j \, d\Gamma^e + \int_{\Gamma_{N_O}^e} N_i (\boldsymbol{q} N_j - \boldsymbol{D} \cdot \nabla N_j) \cdot \boldsymbol{n} \, d\Gamma^e \Big) & \text{divergence form} \\ \sum_e \Big(\int_{\Gamma_C^e} \Phi^e N_i N_j \, d\Gamma^e - \int_{\Gamma_{N_O}^e} N_i (\boldsymbol{D} \cdot \nabla N_j) \cdot \boldsymbol{n} \, d\Gamma^e \Big) & \text{convective form} \end{cases} \\
H_i &= \begin{cases} \sum_e \Big(\int_{\Gamma_C^e} N_i \Phi^{\dagger^e} \phi_C^e \, d\Gamma^e - \int_{\Gamma_N^e \setminus \Gamma_{N_O}^e} N_i q_N^{\dagger^e} \, d\Gamma^e \Big) & \text{divergence form} \\ \sum_e \Big(\int_{\Gamma_C^e} N_i \Phi^e \phi_C^e \, d\Gamma^e - \int_{\Gamma_N^e \setminus \Gamma_{N_O}^e} N_i q_N^e \, d\Gamma^e \Big) & \text{convective form} \end{cases} \\
Q_i &= \sum_e \int_{\Omega^e} N_i H^e \, d\Omega^e - \phi_w Q_w(t)|_i
\end{aligned}$$

[7] Note that:

$$\begin{aligned}
\boldsymbol{q}^e \cdot \nabla N_I^e N_J^e &= \big(N_I^e \boldsymbol{q}^e \cdot \nabla N_J^e \big)^T \\
\boldsymbol{q}^e \cdot \nabla N_i N_j &= \big(N_i \boldsymbol{q}^e \cdot \nabla N_j \big)^T
\end{aligned}$$

$$
\begin{aligned}
&\qquad \boldsymbol{O} = O_{ij} = \sum_e \Big(\sum_I \sum_J O^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
&\boldsymbol{K} = \boldsymbol{A} + \boldsymbol{C} + \boldsymbol{R} + \boldsymbol{B} \\
&\qquad \boldsymbol{A} = A_{ij} = \sum_e \Big(\sum_I \sum_J A^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
&\qquad \boldsymbol{C} = C_{ij} = \sum_e \Big(\sum_I \sum_J C^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
&\qquad \boldsymbol{R} = R_{ij} = \sum_e \Big(\sum_I \sum_J R^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
&\qquad \boldsymbol{B} = B_{ij} = \sum_e \Big(\sum_I \sum_J B^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
&\boldsymbol{F} = \boldsymbol{H} + \boldsymbol{Q} \\
&\qquad \boldsymbol{H} = H_i = \sum_e \Big(\sum_I H^e_I \Delta^e_{Ii} \Big) \\
&\qquad \boldsymbol{Q} = Q_i = \sum_e \Big(\sum_I Q^e_I \Delta^e_{Ii} \Big)
\end{aligned}
\tag{8.103}
$$

with the element matrices

$$
\begin{aligned}
O^e_{IJ} &= \begin{cases} \displaystyle\int_{\Omega^e} \mathcal{R}^e \, N^e_I N^e_J d\Omega^e & \text{divergence form} \\ \displaystyle\int_{\Omega^e} \acute{\mathcal{R}}^e \, N^e_I N^e_J d\Omega^e & \text{convective form} \end{cases} \\
A^e_{IJ} &= \begin{cases} \displaystyle -\int_{\Omega^e} \boldsymbol{q}^e \cdot \nabla N^e_I N^e_J d\Omega^e & \text{divergence form}^7 \\ \displaystyle\int_{\Omega^e} N^e_I \boldsymbol{q}^e \cdot \nabla N^e_J d\Omega^e & \text{convective form} \end{cases} \\
C^e_{IJ} &= \int_{\Omega^e} \nabla N^e_I \cdot (\boldsymbol{D}^e \cdot \nabla N^e_J) d\Omega^e \\
R^e_{IJ} &= \begin{cases} \displaystyle\int_{\Omega^e} (\vartheta^e + \tfrac{\partial \mathcal{R}^e}{\partial t}) N^e_I N^e_J d\Omega^e & \text{divergence form} \\ \displaystyle\int_{\Omega^e} (\vartheta^e + Q^e) N^e_I N^e_J d\Omega^e - \delta_{IJ} Q_w(t)|_I & \text{convective form} \end{cases} \\
B^e_{IJ} &= \begin{cases} \displaystyle\int_{\Gamma^e_C} \Phi^{\dagger^e} N^e_I N^e_J d\Gamma^e + \int_{\Gamma^e_{N_O}} N^e_I (\boldsymbol{q} N^e_J - \boldsymbol{D} \cdot \nabla N^e_J) \cdot \boldsymbol{n} d\Gamma^e & \text{divergence form} \\ \displaystyle\int_{\Gamma^e_C} \Phi^e N^e_I N^e_J d\Gamma^e - \int_{\Gamma^e_{N_O}} N^e_I (\boldsymbol{D} \cdot \nabla N^e_J) \cdot \boldsymbol{n} d\Gamma^e & \text{convective form} \end{cases}
\end{aligned}
\tag{8.104}
$$

and the element vectors

$$
H^e_I = \begin{cases} \displaystyle\int_{\Gamma^e_C} N^e_I \Phi^{\dagger^e} \phi^e_C d\Gamma^e - \int_{\Gamma^e_N \setminus \Gamma^e_{N_O}} N^e_I q^{\dagger^e}_N d\Gamma^e & \text{divergence form} \\ \displaystyle\int_{\Gamma^e_C} N^e_I \Phi^e \phi^e_C d\Gamma^e - \int_{\Gamma^e_N \setminus \Gamma^e_{N_O}} N^e_I q^e_N d\Gamma^e & \text{convective form} \end{cases}
$$

$$Q_I^e = \int_{\Omega^e} N_I^e H^e d\Omega^e - \phi_w Q_w(t)|_I \tag{8.105}$$

where $(i, j = 1, \ldots, N_\text{P})$, $(e = 1, \ldots, N_\text{E})$ and $(I, J = 1, \ldots, N_\text{BN})$. Note that in the $\boldsymbol{B}^e$ matrix we also include the implicit OBC of (8.59) and (8.60) on the specific Neumann boundary $\Gamma^e_{N_O} \subset \Gamma^e_N$. The local element shape functions N_I^e in the element-rank matrices and vectors of (8.104) and (8.105) are expressed in the local $\boldsymbol{\eta}$−coordinate system. Hence, the integration will be done on element level in $\boldsymbol{\eta}$−coordinates. We note in (8.103) and (8.104) that the advection matrix $\boldsymbol{A}$ (and correspondingly $\boldsymbol{A}^e$) as well as the boundary matrix $\boldsymbol{B}$ (and correspondingly $\boldsymbol{B}^e$) at the presence of implicit OBC are unsymmetric, while all other matrices are symmetric.

8.10 Assembly Process

8.10.1 General Procedure

The assembly process is a fundamental feature of finite element computations. Assembly represents the summation of matrix or vector contributions from element integrals to global matrices and vectors. It is mathematically expressed as follows

$$\boldsymbol{K} = K_{ij} = \sum_{e=1}^{N_\text{E}} \Big(\sum_I \sum_J K_{IJ}^e \Delta_{Ii}^e \Delta_{Jj}^e \Big), \quad \boldsymbol{F} = F_i = \sum_{e=1}^{N_\text{E}} \Big(\sum_I F_I^e \Delta_{Ii}^e \Big) \tag{8.106}$$

exemplified for a square matrix $\boldsymbol{A}$ and a column vector $\boldsymbol{F}$. In (8.106) $i, j = 1, \ldots, N_\text{P}$ are the global row and column indices, $I, J = 1, \ldots, N_\text{BN}$ represent local row and column indices, associated with the element matrix K_{IJ}^e and vector F_I^e, and Δ_{Ii}^e and Δ_{Jj}^e are Boolean matrices (8.82) consisting of N_BN rows and N_P columns, which relate the local indices I, J to the global indices i, j. However, in actual computational practice the Boolean matrices Δ_{Ii}^e, Δ_{Jj}^e will never be constructed. Instead, the relation between global and local node numbers is executed via a computer program based on a nodal correspondence table called *incidence matrix*, $\boldsymbol{N} = N_{eJ}$ $(e = 1, \ldots, N_\text{E},\ J = 1, \ldots, N_\text{BN})$. To demonstrate this procedure, we consider an example as shown in Fig. 8.13. A 2D domain is discretized by six linear triangular elements forming a simple eight-noded finite element mesh. The element-node relations are tabulated in the incidence matrix $\boldsymbol{N}$. There is no need to be concerned with the element ordering, however, the assignment of the global node numbers for each element must be consistent and systematic. It is not crucial which first local node of a particular element is incident with one of the global nodes joining the element, however, the remaining global nodes must be counter-clockwise ordered in $\boldsymbol{N}$. This counter-clockwise ordering is consistent with the order of numbering used for the local nodes associated with the master element. For instance, considering element 2 in Fig. 8.13: We have chosen that local node 1 is incident with the global node 2, then local nodes 2 and 3 must be incident with global nodes 4 and 3, respectively. Alternatively, we also could choose for example that the global node 3 is incident with the first local node, so that the global nodes 2 and 4 become incident with the local nodes 2 and 3, respectively. The following C-like pseudo-code explains how the global matrix $\boldsymbol{K}$ and global vector $\boldsymbol{F}$ are assembled from the element matrices $\boldsymbol{K}^e$ and element vectors $\boldsymbol{F}^e$, respectively:

```
$\boldsymbol{K} = \mathbf{0},\ \boldsymbol{F} = \mathbf{0}$                          zeroing global matrix and vector
for ($e = 0; e < N_{\mathrm{E}}; e{+}{+}$) {                 global element loop
   for ($I = 0; I < N_{\mathrm{BN}}; I{+}{+}$) {             local element row loop
      $i = N_{eI}$                                  global row index assignment
      for ($J = 0; J < N_{\mathrm{BN}}; J{+}{+}$) {          local element column loop
         $j = N_{eJ}$                               global column index assignment
         $K_{ij} = K_{ij} + K^e_{IJ}$                     addition of element to the global matrix
      }
      $F_i = F_i + F^e_I$                           addition of element to the global vector
   }
}
```

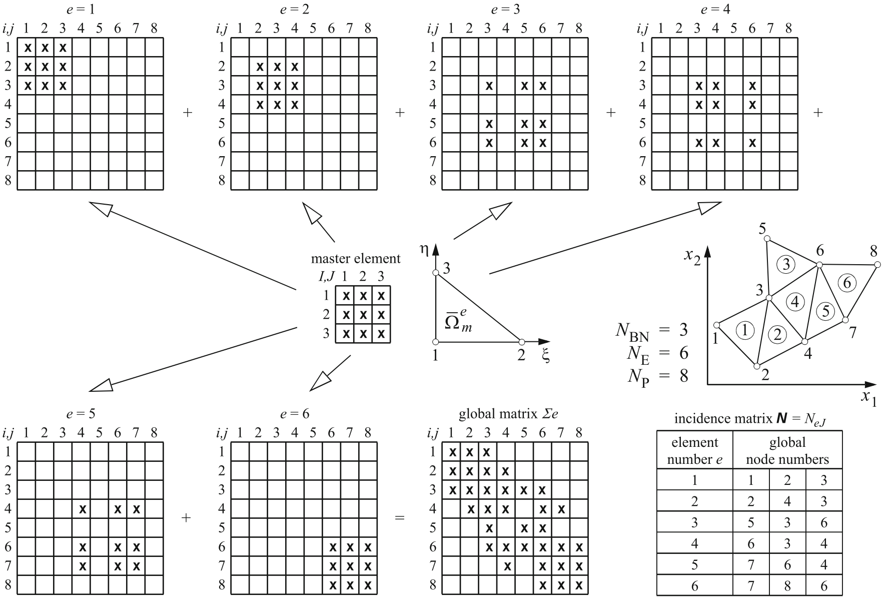
(8.107)

For the example of Fig. 8.13 we obtain finally

$$
K_{ij} = \left(\begin{matrix}
K^1_{11} & K^1_{12} & K^1_{13} & 0 \\
K^1_{21} & K^1_{22} + K^2_{11} & K^1_{23} + K^2_{13} & K^2_{12} \\
K^1_{31} & K^1_{32} + K^2_{31} & K^1_{33} + K^2_{33} + K^3_{22} + K^4_{22} & \\
0 & K^2_{21} & K^2_{23} + K^4_{32} & K^2_{22} + K^4_{33} + K^5_{33} \\
0 & 0 & K^3_{12} & 0 \\
0 & 0 & K^3_{32} + K^4_{12} & K^4_{13} + K^5_{23} \\
0 & 0 & 0 & K^5_{13} \\
0 & 0 & 0 & 0
\end{matrix}\right.
$$

$$
\left.\begin{matrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
K^3_{21} & K^3_{23} + K^4_{21} & 0 & 0 \\
0 & K^4_{31} + K^5_{32} & K^5_{31} & 0 \\
K^3_{11} & K^3_{13} & 0 & 0 \\
K^3_{31} & K^3_{33} + K^4_{11} + K^5_{22} + K^6_{33} & K^5_{21} + K^6_{31} & K^6_{32} \\
0 & K^5_{12} + K^6_{13} & K^5_{11} + K^6_{11} & K^6_{12} \\
0 & K^6_{23} & K^6_{21} & K^6_{22}
\end{matrix}\right) \tag{8.108}
$$

and



incidence matrix $\boldsymbol{N} = N_{eJ}$

element number *e*	global node numbers		
1	1	2	3
2	2	4	3
3	5	3	6
4	6	3	4
5	7	6	4
6	7	8	6

Fig. 8.13 Schematic diagram to illustrate the assembly of a global matrix from element matrices (Modified from [76])

$$F_i = \begin{pmatrix} F_1^1 \\ F_2^1 + F_1^2 \\ F_3^1 + F_3^2 + F_2^3 + F_2^4 \\ F_2^2 + F_3^4 + F_3^5 \\ F_1^3 \\ F_3^3 + F_1^4 + F_2^5 + F_3^6 \\ F_1^5 + F_1^6 \\ F_2^6 \end{pmatrix} \tag{8.109}$$

It can be seen that the resulting global matrix $\boldsymbol{K}$ is sparse and banded. The sparsity structure of $\boldsymbol{K}$ is an advantageous feature of the GFEM. The assembly of the global matrix requires the computation of $N_E\, N_{BN}^2$ coefficients. However, it is not practicable to store the full matrix $\boldsymbol{K}$. In dependence on the preferred equations solvers (see Sect. 8.17.1) different storage management techniques have been developed. For standard iterative equation solvers only the nonzero entries of the matrix are compactly stored, which allows a very core-space saving strategy. Bookkeeping is required to localize the nonzero elements within the matrix. On the other hand, for Gaussian direct equations solvers the profile of the matrix (all elements of a row up to the most right-sided nonzero entry) must be stored, where also zero elements within the matrix profile have to be included, which is needed for the fill-in entries arising at later stages in the Gaussian forward elimination process. While for iterative solvers the order for the global numbering of the nodes is not crucial, the matrix profile used for direct solvers is strongly dependent on the ordering of nodes (and accordingly the demand on storage). As a general rule, the numbering should be such as to minimize the nodal difference for each element (maximum node number minus minimum node number). For large and complex meshes automatic nodal renumbering schemes have to be utilized to minimize the matrix profile (cf. Sect. 8.17.1.4). We note that the order of numbering the elements is neither crucial for iterative nor for direct equation solvers. In cases where the element matrices are symmetric, the global matrix $\boldsymbol{K}$ also becomes symmetric $K_{ij} = K_{ji}$ and only the symmetric half of the matrix coefficients needs to be stored. The assembly of a symmetric global matrix requires the computation of $\frac{1}{2} N_{\mathrm{E}}\, N_{\mathrm{BN}}(N_{\mathrm{BN}} + 1)$ coefficients, which means a reduction of the computational effort by a factor of $(N_{\mathrm{BN}} - 1)/(2N_{\mathrm{BN}})$ compared to an unsymmetric matrix, i.e., for instance $\frac{1}{3}$ and $\frac{5}{12}$ less for 2D triangular and 3D petahedron elements, respectively.

8.10.2 Parallelization via Element Agglomeration

The element-by-element assembly procedure (8.106) and (8.107) in form of $\boldsymbol{K} = \sum_{e}^{N_{\mathrm{E}}} \boldsymbol{K}^{e}$ and so forth is ideally suitable for parallelization by agglomeration of elements for which the element matrices $\boldsymbol{K}^{e}$ can be performed parallel on different CPU's or CPU cores to accelerate significantly the computations. The computational work of assembly is proportional to the number of elements N_{E}. Therefore, an efficient parallel assembly process can be achieved if the total number of elements of a mesh is suitably split into a certain number of subdomains called *partitions* of agglomerated elements so that the element summation is actually executed via

$$\boldsymbol{K} = \sum_{p}^{N_{\mathrm{PA}}} \sum_{e}^{N_{\mathrm{E}p}} \boldsymbol{K}^{e} \quad \text{with} \quad N_{\mathrm{E}} = \sum_{p}^{N_{\mathrm{PA}}} N_{\mathrm{E}_p} \tag{8.110}$$

where N_{PA} is the number of partitions and N_{E_p} is the number of elements agglomerated into partition p. Each partition is concurrently executed on different *threads* representing logical processors. A symmetric multiprocessing facility (SMP) of an operating system distributes all threads to the available physical processors and CPU cores during runtime. However, on computer systems with shared memory to which the threads simultaneously access the summation operations for an element e cannot be executed on different threads at the same time, otherwise the summation becomes erroneous due to occurring *race conditions*. To avoid multiple access during the assembly process the concerned element must be locked while it is summed up. But, *locking* is a rather inefficient process and slows down the computations in particular with increasing number of threads.

More useful and generally preferred in FEFLOW is a technique which is called *disjoint domain partitioning*. It does not need locking and provides an optimal speedup in the parallel assembly process based on (8.110). The computational domain $\bar{\Omega} = \Omega \cup \Gamma$ is subdivided into a maximum number of partitions $\bar{\Omega}^{\mathrm{PAD}}$ of agglomerated elements $\bar{\Omega}^{e}$, which do not join each other, with a remaining (possibly small) border set of partition $\bar{\Omega}^{0}$ (Fig. 8.14), which joins all the disjoint

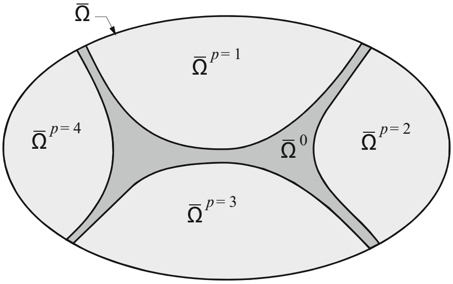


Fig. 8.14 Partitioning of computational domain $\bar{\Omega}$ into disjoint subdomains $\bar{\Omega}^{\text{PAD}} = \bigcup_p^{N_{\text{PAD}}} \bar{\Omega}^p$ (exemplified for $N_{\text{PAD}} = 4$) and border set of partition $\bar{\Omega}^0$ joining $\bar{\Omega}^{\text{PAD}}$

partitions, i.e.,

$$\bar{\Omega} = \bar{\Omega}^{\text{PAD}} \cup \bar{\Omega}^0 = \bigcup_p^{N_{\text{PAD}}} \bar{\Omega}^p \cup \bar{\Omega}^0 \quad \text{with} \quad \bar{\Omega}^p = \bigcup_e^{N_{\text{E}p}} \bar{\Omega}^e \tag{8.111}$$

where a partition p and any other partition q are disjoined if

$$\begin{array}{ll} \bar{\Omega}^p \cap \bar{\Omega}^q = \emptyset \quad \text{with} & \boldsymbol{x}_I^e \in \bar{\Omega}^p \wedge \boldsymbol{x}_I^e \notin \bar{\Omega}^q \\ & (p \neq q, \quad p, q = 1, \ldots, N_{\text{PAD}}) \end{array} \tag{8.112}$$

and

$$\bar{\Omega} \cap \bar{\Omega}^{\text{PAD}} = \bar{\Omega}^0 \neq \emptyset \tag{8.113}$$

for a node I at location $\boldsymbol{x}_I^e$ in element e, where $N_{\text{PAD}} = N_{\text{PA}} - 1$ is the number of disjoint partitions. While all elements belonging to the disjoint partitions can be concurrently assembled in the fast multithreadening mode, the elements of the remaining partition $\bar{\Omega}^0$ must be summed by single threaded execution. However, provided that $\bar{\Omega}^0 \subset \bar{\Omega}$ is small compared to the disjoint partitions $\bar{\Omega}^{\text{PAD}} = \bigcup_p^{N_{\text{PAD}}} \bar{\Omega}^p$, the sequential part of the assembly is insignificant and the parallelized assembly in total provides superior speedups in practical computation. In FEFLOW an efficient and fast agglomeration algorithm is incorporated to find the suitable disjoint partitions of a mesh. To hold $\bar{\Omega}^0$ small as possible, the algorithm runs recursively for cases where $\bar{\Omega}^0$ can be further split into disjoint subpartitions. Practically, no more than three recursions are needed to find the minimum $\bar{\Omega}^0$, which disjoins all $\bar{\Omega}^{\text{PAD}}$ by only one element distance (Fig. 8.15).

8.11 Finite Element Basis Operations

Recalling the basic calculus operations for coordinate transformation described in Sect. 2.1.5, the mapping (8.71) of isoparametric element geometry

$$\boldsymbol{x}^e = \sum_{J=1}^{N_{\text{BN}}} N_J^e(\boldsymbol{\eta})\, \boldsymbol{x}_J^e \tag{8.114}$$

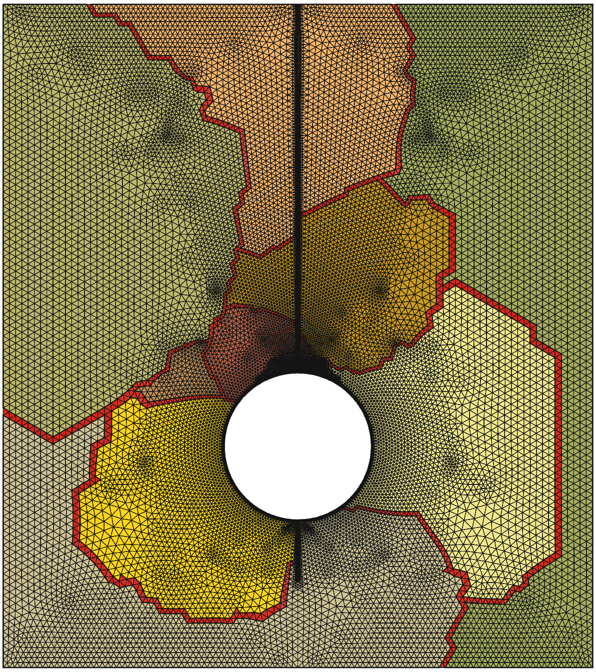


Fig. 8.15 Example of a partitioned 2D triangle mesh: Elements drawn in *red* indicate the border set partition $\bar{\Omega}^0$

between global $\boldsymbol{x}$ and local $\boldsymbol{\eta}$ coordinates is associated with the nonsingular Jacobian $\boldsymbol{J}^e$ defined in the $\Re^3$ space

$$\boldsymbol{J}^e = \frac{\partial \boldsymbol{x}^e}{\partial \boldsymbol{\eta}} = \begin{pmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \\ \frac{\partial}{\partial \zeta} \end{pmatrix} \begin{pmatrix} x_1^e & x_2^e & x_3^e \end{pmatrix} = \begin{pmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{pmatrix}^e = \begin{pmatrix} \frac{\partial x_1^e}{\partial \xi} & \frac{\partial x_2^e}{\partial \xi} & \frac{\partial x_3^e}{\partial \xi} \\ \frac{\partial x_1^e}{\partial \eta} & \frac{\partial x_2^e}{\partial \eta} & \frac{\partial x_3^e}{\partial \eta} \\ \frac{\partial x_1^e}{\partial \zeta} & \frac{\partial x_2^e}{\partial \zeta} & \frac{\partial x_3^e}{\partial \zeta} \end{pmatrix}$$

$$= \begin{pmatrix} \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \xi} x_{1J}^e & \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \xi} x_{2J}^e & \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \xi} x_{3J}^e \\ \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \eta} x_{1J}^e & \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \eta} x_{2J}^e & \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \eta} x_{3J}^e \\ \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \zeta} x_{1J}^e & \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \zeta} x_{2J}^e & \sum\limits_{J=1}^{N_{\mathrm{BN}}} \frac{\partial N_J^e(\xi,\eta,\zeta)}{\partial \zeta} x_{3J}^e \end{pmatrix}, \tag{8.115}$$

in the $\Re^2$ space

$$\boldsymbol{J}^e = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}^e = \begin{pmatrix} \dfrac{\partial x_1^e}{\partial \xi} & \dfrac{\partial x_2^e}{\partial \xi} \\ \dfrac{\partial x_1^e}{\partial \eta} & \dfrac{\partial x_2^e}{\partial \eta} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_{J=1}^{N_{\text{BN}}} \frac{\partial N_J^e(\xi,\eta)}{\partial \xi} x_{1J}^e & \displaystyle\sum_{J=1}^{N_{\text{BN}}} \frac{\partial N_J^e(\xi,\eta)}{\partial \xi} x_{2J}^e \\ \displaystyle\sum_{J=1}^{N_{\text{BN}}} \frac{\partial N_J^e(\xi,\eta)}{\partial \eta} x_{1J}^e & \displaystyle\sum_{J=1}^{N_{\text{BN}}} \frac{\partial N_J^e(\xi,\eta)}{\partial \eta} x_{2J}^e \end{pmatrix} \tag{8.116}$$

and in the $\Re^1$ space

$$\boldsymbol{J}^e = \left(J_{11}\right)^e = \left(\frac{\partial x_1^e}{\partial \xi}\right) = \left(\sum_{J=1}^{N_{\text{BN}}} \frac{\partial N_J^e(\xi)}{\partial \xi} x_{1J}^e\right). \tag{8.117}$$

To evaluate the flux vector divergence terms in the element matrices of (8.104) the inverse Jacobian is required

$$\nabla N_I^e = (\boldsymbol{J}^e)^{-1} \cdot \begin{pmatrix} \frac{\partial N_I^e}{\partial \xi} \\ \frac{\partial N_I^e}{\partial \eta} \\ \frac{\partial N_I^e}{\partial \zeta} \end{pmatrix} \tag{8.118}$$

where

$$(\boldsymbol{J}^e)^{-1} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{x}^e} = \begin{cases} \dfrac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} (J_{22}^e J_{33}^e - J_{32}^e J_{23}^e) & (J_{13}^e J_{32}^e - J_{12}^e J_{33}^e) & (J_{12}^e J_{23}^e - J_{13}^e J_{22}^e) \\ (J_{31}^e J_{23}^e - J_{21}^e J_{33}^e) & (J_{11}^e J_{33}^e - J_{13}^e J_{31}^e) & (J_{21}^e J_{13}^e - J_{23}^e J_{11}^e) \\ (J_{21}^e J_{32}^e - J_{31}^e J_{22}^e) & (J_{12}^e J_{31}^e - J_{32}^e J_{11}^e) & (J_{11}^e J_{22}^e - J_{12}^e J_{21}^e) \end{pmatrix} & \text{in } \Re^3 \\ \dfrac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} J_{22}^e & -J_{12}^e \\ -J_{21}^e & J_{11}^e \end{pmatrix} & \text{in } \Re^2 \\ \dfrac{1}{|\boldsymbol{J}^e|} & \text{in } \Re^1 \end{cases} \tag{8.119}$$

with the determinant of $\boldsymbol{J}^e$

$$|\boldsymbol{J}^e| = \begin{cases} J_{11}^e(J_{22}^e J_{33}^e - J_{32}^e J_{23}^e) - J_{21}^e(J_{12}^e J_{33}^e - J_{13}^e J_{32}^e) + J_{31}^e(J_{12}^e J_{23}^e - J_{13}^e J_{22}^e) & \text{in } \Re^3 \\ J_{11}^e J_{22}^e - J_{21}^e J_{12}^e & \text{in } \Re^2 \\ J_{11}^e & \text{in } \Re^1 \end{cases} \tag{8.120}$$

Suppose the local shape functions are continuous and at least once-differentiable with respect to the local coordinates $\boldsymbol{\eta}$, a necessary and sufficient condition for $(\boldsymbol{J}^e)^{-1}$ to exist is that the determinant of the Jacobian $|\boldsymbol{J}^e|$ of element e be nonzero at every point $\boldsymbol{\eta}$ in $\bar{\Omega}^e$:

$$|\boldsymbol{J}^e| \neq 0 \tag{8.121}$$

The master element matrices and vectors appearing in (8.104) and (8.105), respectively, are to be integrated over element volumes Ω^e and surfaces Γ^e. The integration in local coordinates becomes for a differential 'volume' element

$$d\Omega^e = \left\{ \begin{array}{lll} \left.\begin{array}{ll} dx_1\, dx_2\, dx_3 &= |\boldsymbol{J}^e| d\xi\, d\eta\, d\zeta \\ dx_1\, dx_2 &= |\boldsymbol{J}^e| d\xi\, d\eta \\ dx_1 &= |\boldsymbol{J}^e| d\xi \end{array}\right\} & \text{Cartesian} & \Re^D \; (D = 1, 2, 3) \\ rdrd\phi dz \;= 2\pi |\boldsymbol{J}^e| rd\xi\, d\eta & \text{axisymmetric} & \end{array} \right. \tag{8.122}$$

and for a differential 'areal' element in Cartesian coordinates of $\Re^D$ $(D = 1, 2, 3)$ space:

$$d\Gamma^e = \left\{ \begin{array}{llll}
\left\| \begin{pmatrix} \frac{\partial x_1}{\partial \xi} \\ \frac{\partial x_2}{\partial \xi} \\ \frac{\partial x_3}{\partial \xi} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x_1}{\partial \eta} \\ \frac{\partial x_2}{\partial \eta} \\ \frac{\partial x_3}{\partial \eta} \end{pmatrix} \right\| d\xi d\eta & = \left\| \det \begin{pmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \boldsymbol{e}_3 \\ J^e_{11} & J^e_{12} & J^e_{13} \\ J^e_{21} & J^e_{22} & J^e_{23} \end{pmatrix} \right\| d\xi d\eta & & \\
& = \left\| \begin{pmatrix} J^e_{12}J^e_{23} - J^e_{13}J^e_{22} \\ J^e_{13}J^e_{21} - J^e_{11}J^e_{23} \\ J^e_{11}J^e_{22} - J^e_{12}J^e_{21} \end{pmatrix} \right\| d\xi d\eta & \text{at } \zeta = \pm 1 & \text{in } \Re^3 \\
\left\| \begin{pmatrix} \frac{\partial x_1}{\partial \eta} \\ \frac{\partial x_2}{\partial \eta} \\ \frac{\partial x_3}{\partial \eta} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x_1}{\partial \zeta} \\ \frac{\partial x_2}{\partial \zeta} \\ \frac{\partial x_3}{\partial \zeta} \end{pmatrix} \right\| d\eta d\zeta & = \left\| \det \begin{pmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \boldsymbol{e}_3 \\ J^e_{21} & J^e_{22} & J^e_{23} \\ J^e_{31} & J^e_{32} & J^e_{33} \end{pmatrix} \right\| d\eta d\zeta & & \\
& = \left\| \begin{pmatrix} J^e_{22}J^e_{33} - J^e_{23}J^e_{32} \\ J^e_{23}J^e_{31} - J^e_{21}J^e_{33} \\ J^e_{21}J^e_{32} - J^e_{22}J^e_{31} \end{pmatrix} \right\| d\eta d\zeta & \text{at } \xi = \pm 1 & \text{in } \Re^3 \\
\left\| \begin{pmatrix} \frac{\partial x_1}{\partial \xi} \\ \frac{\partial x_2}{\partial \xi} \\ \frac{\partial x_3}{\partial \xi} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x_1}{\partial \zeta} \\ \frac{\partial x_2}{\partial \zeta} \\ \frac{\partial x_3}{\partial \zeta} \end{pmatrix} \right\| d\xi d\zeta & = \left\| \det \begin{pmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \boldsymbol{e}_3 \\ J^e_{11} & J^e_{12} & J^e_{13} \\ J^e_{31} & J^e_{32} & J^e_{33} \end{pmatrix} \right\| d\xi d\zeta & & \\
& = \left\| \begin{pmatrix} J^e_{12}J^e_{33} - J^e_{13}J^e_{32} \\ J^e_{13}J^e_{31} - J^e_{11}J^e_{33} \\ J^e_{11}J^e_{32} - J^e_{12}J^e_{31} \end{pmatrix} \right\| d\xi d\zeta & \text{at } \eta = \pm 1 & \text{in } \Re^3 \\
\left\| \begin{pmatrix} \frac{\partial x_1}{\partial \xi} \\ \frac{\partial x_2}{\partial \xi} \end{pmatrix} \right\| d\xi = \left\| \begin{pmatrix} J^e_{11} \\ J^e_{12} \end{pmatrix} \right\| d\xi & = \sqrt{(J^e_{11})^2 + (J^e_{12})^2} d\xi & \text{at } \eta = \pm 1 & \text{in } \Re^2 \\
\left\| \begin{pmatrix} \frac{\partial x_1}{\partial \eta} \\ \frac{\partial x_2}{\partial \eta} \end{pmatrix} \right\| d\eta = \left\| \begin{pmatrix} J^e_{21} \\ J^e_{22} \end{pmatrix} \right\| d\eta & = \sqrt{(J^e_{21})^2 + (J^e_{22})^2} d\eta & \text{at } \xi = \pm 1 & \text{in } \Re^2 \\
\ldots |_{\xi=-1}^{\xi=+1} \quad \text{in} \quad \Re^1 & & &
\end{array} \right. \tag{8.123}$$

and in cylindrical coordinates of $\Re^2$ (meridional) space (Fig. 8.16):

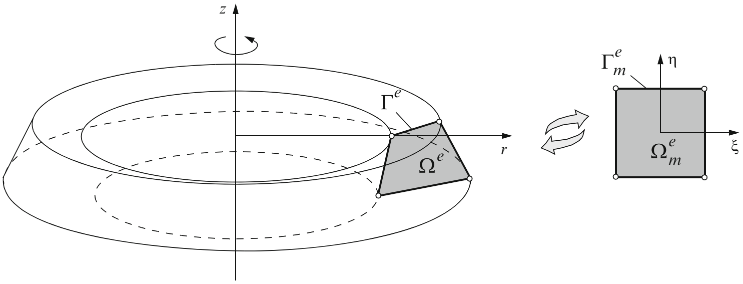


Fig. 8.16 Axisymmetric finite element in global (*cylindrical*) and local coordinates

$$d\Gamma^e = \begin{cases} \left\| \begin{pmatrix} \frac{\partial r}{\partial \xi} \\ \frac{\partial z}{\partial \xi} \end{pmatrix} \right\| r d\xi d\phi = 2\pi \left\| \begin{pmatrix} J_{11}^e \\ J_{12}^e \end{pmatrix} \right\| r d\xi = 2\pi \sqrt{(J_{11}^e)^2 + (J_{12}^e)^2} r d\xi \quad \text{at } \eta = \pm 1 \\ \left\| \begin{pmatrix} \frac{\partial r}{\partial \eta} \\ \frac{\partial z}{\partial \eta} \end{pmatrix} \right\| r d\eta d\phi = 2\pi \left\| \begin{pmatrix} J_{21}^e \\ J_{22}^e \end{pmatrix} \right\| r d\eta = 2\pi \sqrt{(J_{21}^e)^2 + (J_{22}^e)^2} r d\eta \quad \text{at } \xi = \pm 1 \end{cases} \tag{8.124}$$

where $r = \sum_J^{N_{\text{BN}}} N_J^e(\xi, \eta) r_J^e$.

In using (8.114)–(8.124) all integrals in (8.104) and (8.105) can be expressed in the $\boldsymbol{\eta}$–coordinate system. For example, the element advection matrix $\boldsymbol{A}^e$ of the convective form (8.104) becomes in 3D

$$\begin{aligned} A_{IJ}^e &= \int_{\Omega^e} N_I^e \boldsymbol{q} \cdot \nabla N_J^e d\Omega^e \\ &= \iiint_{-1}^{+1} N_I^e(\boldsymbol{\eta}) \boldsymbol{q} \cdot \left((\boldsymbol{J}^e)^{-1} \cdot \begin{pmatrix} \frac{\partial N_J^e(\boldsymbol{\eta})}{\partial \xi} \\ \frac{\partial N_J^e(\boldsymbol{\eta})}{\partial \eta} \\ \frac{\partial N_J^e(\boldsymbol{\eta})}{\partial \zeta} \end{pmatrix} \right) |\boldsymbol{J}^e| d\xi d\eta d\zeta \end{aligned} \tag{8.125}$$

and similarly for all other element integrals appearing in (8.104) and (8.105). Typically, the integrals always take the form[8]:

[8] The local coordinates $\boldsymbol{\eta}$ commonly range $-1 \le \boldsymbol{\eta} \le +1$, except in triangular or tetrahedral (and partly pentahedral and pyramidal) geometries, where the lower limit is zero: $0 \le \boldsymbol{\eta} \le +1$, see furthermore (8.128).

$$
\begin{aligned}
\textstyle\int_{\Omega^e} f(\boldsymbol{x})d\Omega^e &= \begin{cases} \displaystyle\iiint_{-1}^{+1} f(\xi,\eta,\zeta)d\xi d\eta d\zeta & \text{in } \Re^3 \\ \displaystyle\iint_{-1}^{+1} f(\xi,\eta)d\xi d\eta & \text{in } \Re^2 \\ \displaystyle\int_{-1}^{+1} f(\xi)d\xi & \text{in } \Re^1 \end{cases} \\
\textstyle\int_{\Gamma^e} g(\boldsymbol{x})d\Gamma^e &= \begin{cases} \displaystyle\iint_{-1}^{+1} g(\xi,\eta,\zeta)(d\xi d\eta, d\eta d\zeta, d\xi d\zeta) & \text{in } \Re^3 \\ \displaystyle\int_{-1}^{+1} g(\xi,\eta)(d\xi, d\eta) & \text{in } \Re^2 \\ g(\xi)|_{\xi=-1}^{\xi=+1} & \text{in } \Re^1 \end{cases}
\end{aligned}
\tag{8.126}
$$

where $f(.)$ and $g(.)$ are volume and surface integrand functions, respectively. We observe that the dependency of the element e on the physical (global) geometry only occurs in the Jacobian $\boldsymbol{J}^e$. Since the coordinate transformation is relatively simple, the Jacobian matrix should be easily evaluated.

8.12 Numerical and Analytical Integration

The transformation of the geometry and the variable coefficients in the element integrals (8.104) and (8.105) from the global coordinates $\boldsymbol{x}$ to the local coordinates $\boldsymbol{\eta}$ results in algebraically complex expressions, which cannot be analytically evaluated for distorted element geometries in general. However, in fact this is not an intrinsic disadvantage because very efficient and *exact* numerical integration techniques are available which makes the master element integration very cost-effective and highly flexible for a wide class of finite elements under general geometric (i.e., distorted) conditions.

For our needs the most efficient and, therefore, preferable numerical integration is the *Gauss-Legendre quadrature*, e.g., [590], providing an optimal degree of precision. Thus, the evaluation of the 3D, 2D and 1D elemental integrals of (8.126) reduces to expressions with a triple, double and single summation, respectively, in the following form:

$$
\begin{aligned}
\iiint_{-1}^{+1} f(\xi,\eta,\zeta)d\xi d\eta d\zeta &= \sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n} H_p H_q H_r f(\xi_p,\eta_q,\zeta_r) \\
\iint_{-1}^{+1} f(\xi,\eta)d\xi d\eta &= \sum_{p=1}^{n}\sum_{q=1}^{n} H_p H_q f(\xi_p,\eta_q) \\
\int_{-1}^{+1} f(\xi)d\xi &= \sum_{p=1}^{n} H_p f(\xi_p)
\end{aligned}
\tag{8.127}
$$

Table 8.2 Gauss-Legendre quadrature sampling points and weights for $\int_{-1}^{+1} f(\xi)d\xi = \sum_{p=1}^{n} H_p f(\xi_p)$

Order	n	p	ξ_p	H_p	
Linear	1	1	0	2	
Quadratic	2	1	$+1/\sqrt{3}$	1	
		2	$-1/\sqrt{3}$	1	
Cubic	3	1	$+\sqrt{0.6}$	$5/9$	
		2	0	$8/9$	
		3	$-\sqrt{0.6}$	$5/9$	
Quintic	4	1	$+\sqrt{(3+a)/7}$	$0.5 - 1/(3a)$	$a = \sqrt{4.8}$
		2	$+\sqrt{(3-a)/7}$	$0.5 + 1/(3a)$	
		3	$-\sqrt{(3-a)/7}$	$0.5 + 1/(3a)$	
		4	$-\sqrt{(3+a)/7}$	$0.5 - 1/(3a)$	

where $n \geq 1$ is the number of quadrature points in each direction, ξ_p, η_q, ζ_r denote the Gauss point local coordinates in $\bar{\Omega}_m^e$ and H_p, H_q, H_r are the associated quadrature rule weights. Note that for 1D elements the integrals are commonly evaluated analytically. Choosing n Gauss points a polynomial expression $f(.)$ of degree $2n - 1$ can be exactly integrated. The positions and weights for the Gauss-Legendre quadrature rule up to order $n = 4$ are listed in Table 8.2. When the integrand $f(.)$ is of different degree in ξ, η, ζ, the number of Gauss points should be selected on the basis of the largest-degree polynomial. The minimum allowable quadrature is one that yields the volume or area of the element exactly [590]. For undistorted elements (such as rectangular or brick-shaped elements) the 2- and 3-point Gauss-Legendre rules (i.e., Gauss points in each direction) are sufficient to evaluate exactly all interesting integrals of linear and quadratic element types, respectively, because the Jacobian of the mapping $\boldsymbol{J}^e$ is constant for these geometries. However, for distorted elements the Jacobian is no more constant and integrals involving more than one derivative cannot be exactly integrated since the integrand is a quotient of two polynomials [341]. In general, $f(.)$ may not be really polynomial due to the complex dependency of the integrand on $\boldsymbol{J}^e$ and the possible presence of other variable coefficients such that the required number of Gauss points can only be estimated. In practice, an *optimal order of Gauss-Legendre integration* is used, which is defined as one that guarantees the highest possible accuracy and rate of convergence while minimizing the computational cost. For linear quadrilateral elements $2 \times 2 \times 2$ and 2×2, and for quadratic quadrilateral elements $3 \times 3 \times 3$ and 3×3 are of optimal order in 3D and 2D, respectively. For the areal triangular element (triangle), the 3D linear tetrahedron, the 3D linear triangular prism (petahedron) and the 3D linear pyramidal element (pyramid) specific integration points and weights are applied [84, 280, 586, 590] as follows:

Table 8.3 Quadrature points and weights of formulae (8.128) of quadratic order for linear triangle, linear tetrahedron, linear pentahedron and linear square pyramid

Type	m	p	ξ_p	η_p	ζ_p	H_p	
Triangle	3	1	1/2	0		1/6	
		2	1/2	1/2		1/6	
		3	0	1/2		1/6	
Tetrahedron	4	1	a	a	a	1/24	$a = 0.13819660$
		2	b	a	a	1/24	$b = 0.58541020$
		3	a	b	a	1/24	
		4	a	a	b	1/24	
Pentahedron	6	1	1/2	0	$+1/\sqrt{3}$	1/6	
		2	1/2	1/2	$+1/\sqrt{3}$	1/6	
		3	0	1/2	$+1/\sqrt{3}$	1/6	
		4	1/2	0	$-1/\sqrt{3}$	1/6	
		5	1/2	1/2	$-1/\sqrt{3}$	1/6	
		6	0	1/2	$-1/\sqrt{3}$	1/6	
Pyramid	8	1	$+c/\sqrt{3}$	$+c/\sqrt{3}$	$1-c$	0.1007858820798250	$c = 0.455848155988775$
		2	$-c/\sqrt{3}$	$+c/\sqrt{3}$	$1-c$	0.1007858820798250	
		3	$+c/\sqrt{3}$	$-c/\sqrt{3}$	$1-c$	0.1007858820798250	
		4	$-c/\sqrt{3}$	$-c/\sqrt{3}$	$1-c$	0.1007858820798250	
		5	$+d/\sqrt{3}$	$+d/\sqrt{3}$	$1-d$	0.2325474512535080	$d = 0.877485177344559$
		6	$-d/\sqrt{3}$	$+d/\sqrt{3}$	$1-d$	0.2325474512535080	
		7	$+d/\sqrt{3}$	$-d/\sqrt{3}$	$1-d$	0.2325474512535080	
		8	$-d/\sqrt{3}$	$-d/\sqrt{3}$	$1-d$	0.2325474512535080	

$$
\begin{aligned}
&\text{triangle}\\
&\int_{\Gamma^e} f(\boldsymbol{x})d\Gamma^e = \int_0^1\int_0^{1-\xi} f(\xi,\eta)d\eta d\xi = \sum_{p=1}^{m} H_p f(\xi_p,\eta_p)\\
&\text{tetrahedron}\\
&\int_{\Omega^e} f(\boldsymbol{x})d\Omega^e = \int_0^1\int_0^{1-\xi}\int_0^{1-\xi-\eta} f(\xi,\eta,\zeta)d\zeta d\eta d\xi = \sum_{p=1}^{m} H_p f(\xi_p,\eta_p,\zeta_p)\\
&\text{pentahedron}\\
&\int_{\Omega^e} f(\boldsymbol{x})d\Omega^e = \int_0^1\int_0^{1-\xi}\int_{-1}^{1} f(\xi,\eta,\zeta)d\zeta d\eta d\xi = \sum_{p=1}^{m} H_p f(\xi_p,\eta_p,\zeta_p)\\
&\text{pyramid}\\
&\int_{\Omega^e} f(\boldsymbol{x})d\Omega^e = \int_0^1\int_{\zeta-1}^{1-\zeta}\int_{\zeta-1}^{1-\zeta} f(\xi,\eta,\zeta)d\zeta d\eta d\xi = \sum_{p=1}^{m} H_p f(\xi_p,\eta_p,\zeta_p)
\end{aligned}
\tag{8.128}
$$

where m is the total number of integration points. The sampling points and weights for the linear triangle, linear tetrahedron, linear pentahedron and linear square pyramid of quadratic order with $m = 3$, $m = 4$, $m = 6$ and $m = 8$, respectively, are listed in Table 8.3. Figure 8.17 illustrates the locations of the integration points

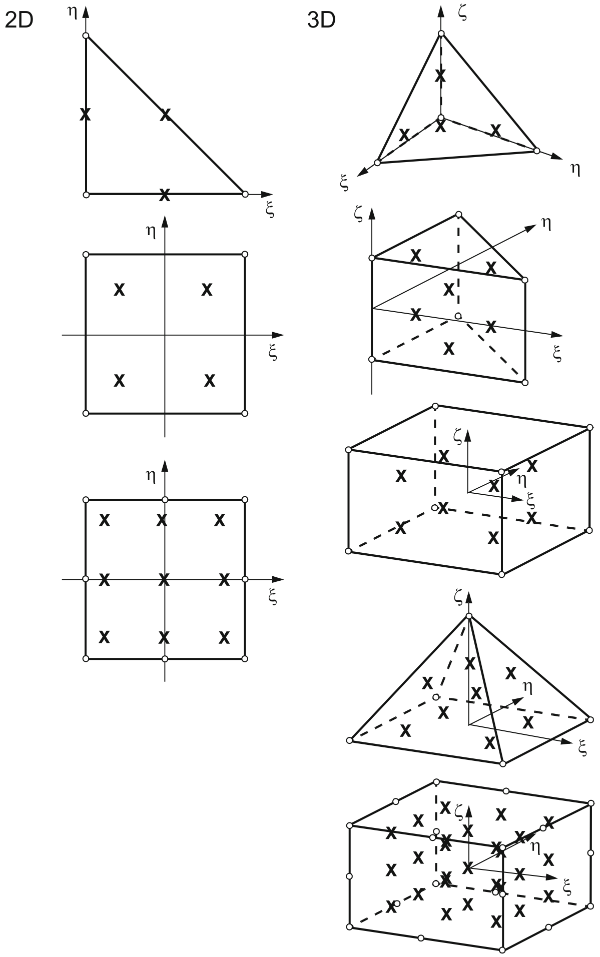
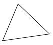
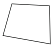
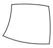
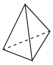
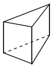
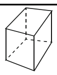
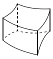


Fig. 8.17 Location of Gauss points symbolized by × for linear and quadratic isoparametric elements used in FEFLOW in 2D (3-point rule for linear triangle, 4-point rule for linear quadrilateral and 9-point rule for quadratic quadrilateral) and 3D (4-point rule for linear tetrahedron, 6-point rule linear pentahedron, 8-point rule for linear hexahedron, 8-point rule for linear pyramid, 27-point rule for quadratic hexahedron)

in the master element $\bar{\Omega}_m^e$ for the 2D and 3D finite elements used in FEFLOW. For axisymmetric elements the same quadrature rules are applied than in 2D.

Table 8.4 gives an estimation of arithmetic operations required for the numerical integration of one element. The working steps 1–7 in Table 8.4 indicate the effort from the coordinate transformation, which has to be performed basically, however, only once for all integrands in consideration. We recognize that the computation of the Jacobian and the terms related to the global derivatives are the most expensive steps. Once these terms are available the effort in computing additional terms and extensions in the integrand functions (e.g., introducing variable coefficients and anisotropic relations) remains low, which makes the numerical integration very flexible and efficient. We also see that the required number of total operations increases significantly with more complex (higher-order) finite elements, in particular for 3D finite elements, however, in favor of a higher accuracy in the

Table 8.4 Estimation of computational effort for numerical integration of 2D and 3D finite elements: Number of required arithmetic operations (sum of multiplications, divisions, additions and subtractions) to build typical matrices of (8.104) for one element e via Gauss-Legendre quadrature

		2D			3D			
	Element type							
		3-node triangle	4-node quadrilateral	8-node quadrilateral	4-node tetrahedron	6-node pentahedron	8-node hexahedron	20-node hexahedron
	N_{BN}	3	4	8	4	6	8	20
	Gauss points m	3	4	9	4	6	8	27
1	N_I^e	2	17	48	3	22	68	255
2	$\frac{\partial N_I^e}{\partial \xi}$	0	4	22	0	4	16	112
3	$\frac{\partial N_I^e}{\partial \eta}$	0	4	22	0	4	16	112
4	$\frac{\partial N_I^e}{\partial \zeta}$	0	0	0	0	5	16	128
5	$\boldsymbol{J}^e$ $(\propto\ 2\,D^2\,N_{\mathrm{BN}})$	24	32	64	72	108	144	360
6	$\lvert\boldsymbol{J}^e\rvert\,d\Omega$	4	4	4	15	15	15	15
7	$(\boldsymbol{J}^e)^{-1}$	5	5	5	28	28	28	28
8	∇N_I^e $(\propto\ 2\,D^2\,N_{\mathrm{BN}})$	24	32	64	72	108	144	360
9	$\int N_I^e$ $(\propto\ N_{\mathrm{BN}})$	3	4	8	4	6	8	20
10	$\int N_I^e N_J^e$ $(\propto\ \frac{1}{2} N_{\mathrm{BN}}(N_{\mathrm{BN}}+1))$	6	10	36	10	21	36	210
11	$\int N_I^e \boldsymbol{q}\cdot\nabla N_J^e$ $(\propto\ 2\,D\,N_{\mathrm{BN}}^2)$	36	64	256	96	216	384	2,400
12	$\int \nabla N_I^e\cdot(\boldsymbol{D}\cdot\nabla N_J^e)$ $(\propto\ D^2\,N_{\mathrm{BN}}(N_{\mathrm{BN}}+1))$	48	80	288	180	378	648	3,780
Operations per Gauss point (summation of steps 1–12)		152	256	817	480	915	1,523	7,780
Operations per element (multiplied by m)		456	1,024	7,353	1,920	5,490	12,184	210,060

approximation. Nevertheless, to maintain the generality in the geometric shapes of the used finite elements the numerical integration is indispensable. All quadrilateral, pentahedral and hexahedral elements lead to very complicated integral expression which can only be tackled numerically. Analytical evaluation is available only for specific element shapes. In Appendix H we evaluate the element matrices of (8.104) on an analytical basis for the linear 1D element, the linear 2D triangle and the linear 3D tetrahedron assuming constant coefficients. Only for these types of elements the Jacobians are always constant for every element shape, which make these elements favorable to analytical integration. In Appendix H we also discuss exceptions for the quadrilateral, hexahedral, pentahedral and pyramidal element, where a constant Jacobian is only attainable for an undisturbed element shape, such as the rectangle or parallelogram for the quadrilateral element, the brick or parallelepiped for the hexahedral element, the triangular prism with parallel top and bottom surfaces for the pentahedral element and the pyramid with a parallelogram or rectangular base and oblique shape for the pyramidal element.

8.13 Temporal Discretization

8.13.1 General

The semidiscrete Galerkin approximation (8.96) of the governing ADE has led to system of ordinary differential equations (ODE's) written in the form (8.100)

$$\boldsymbol{O} \cdot \dot{\boldsymbol{\phi}} + \boldsymbol{K} \cdot \boldsymbol{\phi} = \boldsymbol{F} \tag{8.129}$$

for solving $\boldsymbol{\phi} = \boldsymbol{\phi}(t)$ associated with IC's at $t = t_0$

$$\boldsymbol{\phi}(t_0) = \boldsymbol{\phi}_0 \tag{8.130}$$

where $\boldsymbol{O}$ is called a *consistent mass* (CM) matrix because it is defined consistent with the weak formulation in (8.98) and (8.99) assuming the separability of space $\boldsymbol{x}$ and time t. It remains to solve the resulting semidiscrete equations (8.129) for $\boldsymbol{\phi}(t)$ via appropriate and cost-effective time-integration methods, which integrate (8.129) in time t to trace the temporal evolution of $\boldsymbol{\phi}(t)$ from the initial solution $\boldsymbol{\phi}_0$.

Let us rewrite (8.129) in a normalized form, viz.,

$$\dot{\boldsymbol{\phi}} + \boldsymbol{\mu} \cdot \boldsymbol{\phi} = \boldsymbol{f} \quad \text{with} \quad \boldsymbol{\mu} = \boldsymbol{O}^{-1} \cdot \boldsymbol{K} \quad \text{and} \quad \boldsymbol{f} = \boldsymbol{O}^{-1} \cdot \boldsymbol{F} \tag{8.131}$$

provided that $\boldsymbol{O}$ is invertible with $|\boldsymbol{O}| \neq 0$, we find the solution of this first-order system of ODE's as [10]

$$\boldsymbol{\phi}(t) = \underbrace{e^{-\boldsymbol{\mu}(t-t_0)} \cdot \boldsymbol{\phi}_0}_{\text{decay}} + \int_{t_0}^{t} \underbrace{e^{-\boldsymbol{\mu}(t-\tau)} \cdot \boldsymbol{f}(\tau)}_{\text{forcing}} \, d\tau \tag{8.132}$$

consisting of two components: (1) the exponential decay of the homogeneous part and (2) the particular solution of the forcing contribution. However, we recognize that the exponential matrix $e^{-\boldsymbol{\mu}t}$ is complex and not an algebraic statement, hence not directly solvable. We have to conclude that, in general, it is not possible to integrate (8.132) (and accordingly (8.129)) on an analytical basis and further approximation methods are required to obtain a set of algebraic equations in terms of the nodal state-variable $\boldsymbol{\phi}$.

There is an abundance of numerical methods for solving ODE's, which are categorized as *methods of lines* in the classic literature [331, 441, 462]. Among the wide variety of available methods, however, from the practical point of view, in particular the computational cost, we have interests only in efficient two-stage *single-step* time marching *recurrence* schemes, where for stability reasons implicit

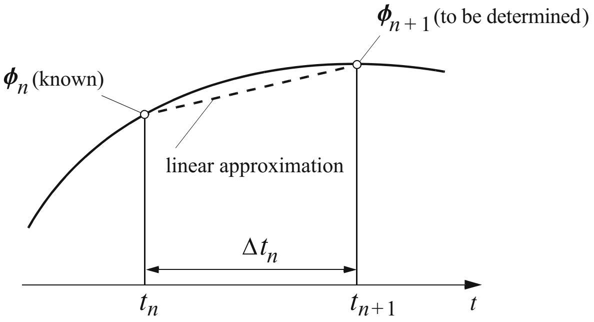


Fig. 8.18 Time marching recurrence of ϕ in the finite time interval (t_n, t_{n+1}), where $t_{n+1} = t_n + \Delta t_n$

or semi-implicit (so-called $\mathcal{A}$–stable [110]) algorithms will be preferred, which are stable independent of the used time step. Considering $\phi(t)$ within the finite interval (t_n, t_{n+1}) with

$$t_{n+1} = t_n + \Delta t_n \tag{8.133}$$

where the subscript n denotes the time plane and $\Delta t_n = t_{n+1} - t_n$ is a variable time step length, the state-variable $\phi(t)$ is defined as

$$\phi_n = \phi(t_n) \tag{8.134}$$

at the previous (old) time plane n and as

$$\phi_{n+1} = \phi(t_{n+1}) \tag{8.135}$$

at the new time plane $n + 1$. In each interval, ϕ_{n+1} is recursively solved from the preceding values ϕ_n at beginning of the time step Δt_n as shown in Fig. 8.18, and (8.132) can be recast into an incremental form, viz.,

$$\phi_{n+1} = e^{-\boldsymbol{\mu}\Delta t_n} \cdot \phi_n + e^{-\boldsymbol{\mu} t_{n+1}} \cdot \int_{t_n}^{t_{n+1}} e^{\boldsymbol{\mu}\tau} \cdot \boldsymbol{f}(\tau) d\tau \tag{8.136}$$

While (8.136) is still an exact solution in time t without any approximation, it is necessary to expand the exponential decay matrix $e^{-\boldsymbol{\mu}\Delta t_n}$ within the time step Δt_n into a power series. The most typical linear approximations for $e^{-\boldsymbol{\mu}\Delta t_n}$ are listed in Table 8.5 and will be discussed in the following. However, before we proceed and introduce the appropriate time stepping schemes in detail, we have to consider first the approximation of the mass matrix $\boldsymbol{O}$ called *mass lumping*.

Table 8.5 Linear $\theta-$approximations[a] of the exponential decay matrix $e^{-\boldsymbol{\mu}\Delta t_n}$

Algorithm	θ	Approximation
Forward	0	$\boldsymbol{\delta} - \Delta t_n \boldsymbol{\mu}$
Backward	1	$[\boldsymbol{\delta} + \Delta t_n \boldsymbol{\mu}]^{-1}$
Trapezoid	$\frac{1}{2}$	$[\boldsymbol{\delta} + (\Delta t_n/2)\boldsymbol{\mu}]^{-1} \cdot [\boldsymbol{\delta} - (\Delta t_n/2)\boldsymbol{\mu}]$
Galerkin	$\frac{2}{3}$	$[\boldsymbol{\delta} + (2\Delta t_n/3)\boldsymbol{\mu}]^{-1} \cdot [\boldsymbol{\delta} - (\Delta t_n/3)\boldsymbol{\mu}]$

[a] Weighting coefficient ($0 \leq \theta \leq 1$) classifies approximation methods

8.13.2 Mass Lumping

The Galerkin formulation naturally leads to consistent mass matrices $\boldsymbol{O} = \sum_e \boldsymbol{O}^e$, which typically distribute the mass of an element over all associated nodes. This can be seen in the discrete element mass matrices $\boldsymbol{O}^e$ of (H.8), (H.23) and (H.41) in Appendix H for the linear 1D, 2D triangular and 3D tetrahedral element, respectively. However, there can be different numerical reasons to concentrate (lump) the mass of an element on the mesh nodes. Mass lumping is a typical feature of the FDM and can also be useful in the finite element context for certain time stepping schemes. The principal motivation behind this technique is the generation of a mass matrix $\boldsymbol{O}$, which is *diagonal* and readily invertible to evaluate $\boldsymbol{O}^{-1}$ of (8.131) in a trivial way. In contrast, the CM matrix, which is sparse and banded, has an inverse that is dense, such that the formulation (8.131) becomes inferior to (8.129) in practical computations.

To permit an equivalent formulation of the mass matrix $\boldsymbol{O}^e$ of an element e we replace

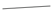
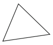
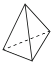
$$\boldsymbol{O}^e = O^e_{IJ} = \underbrace{\int_{\Omega^e} \acute{\mathcal{R}}^e N^e_I N^e_J d\Omega^e}_{\text{consistent}} \rightarrow \underbrace{\delta_{IJ} \int_{\Omega^e} \acute{\mathcal{R}}^e N^e_I d\Omega^e}_{\text{lumping}} \tag{8.137}$$

where δ_{IJ} is the Kronecker symbol (2.7). Since $\sum_{J=1}^{N_{\rm BN}} N^e_J = 1$, (8.67), the lumping procedure is equivalent to *summing* the rows of the CM matrix: $O^e_{II} = \sum_J O^e_{IJ}$. However, this row-summing technique of mass lumping is usually only applicable to linear elements, where the diagonals are always positive. We note that for higher-order elements specific rules of mass lumping are required [590], for instance [129]:

$$O^e_{IJ} = \begin{cases} \int_{\Omega^e} \acute{\mathcal{R}}^e d\Omega^e \int_{\Omega^e} N^e_I N^e_J d\Omega^e \Big/ \sum_{I=1}^{N_{\rm BN}} \int_{\Omega^e} N^e_I N^e_J d\Omega^e & \text{for} \quad I = J, \\ 0 & \text{for} \quad I \neq J \end{cases} \tag{8.138}$$

In Table 8.6 analytical formulations of the consistent mass (CM) and the lumped mass (LM) matrix are compared for the linear 1D, 2D triangular and 3D tetrahedral element (cf. Appendix H). Using numerical integration, the quadrature rules are directly applied to the $\delta_{IJ} \int_{\Omega^e} N^e_I d\Omega^e$ term to yield a diagonal matrix for $\boldsymbol{O}^e$.

Table 8.6 Consistent versus lumped mass matrix for the linear 1D, 2D triangular and 3D tetrahedral element (cf. Appendix H)

	Consistent	Lumping
Element type	$\int_{\Omega^e} N_I^e N_J^e d\Omega^e$	$\delta_{IJ} \int_{\Omega^e} N_I^e d\Omega^e$
	$\frac{\Delta x^e}{6}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}$	$\frac{\Delta x^e}{2}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$
	$\frac{A^e}{12}\begin{pmatrix}2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2\end{pmatrix}$	$\frac{A^e}{3}\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$
	$\frac{V^e}{20}\begin{pmatrix}2 & 1 & 1 & 1\\ 1 & 2 & 1 & 1\\ 1 & 1 & 2 & 1\\ 1 & 1 & 1 & 2\end{pmatrix}$	$\frac{V^e}{4}\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}$

Mass lumping is generally desired for an explicit time integration scheme to perform $\boldsymbol{O}^{-1}$ very easily in use of (8.131) because the application of a CM matrix for the explicit scheme becomes too expensive. On the other hand, for implicit time integration methods mass lumping is commonly neither necessary nor preferred. In general, CM matrix formulations provide higher accuracy, see [209, 210, 590]. Exception is given for unsaturated flow problems, where fully implicit time stepping schemes in combination with mass lumping have shown superior to CM.

8.13.3 Galerkin Approximation in Time

Similar to the spatial approximation (8.16) we can use the standard finite element expansion for the time-dependent state variable as

$$\phi(t) \approx \hat{\phi}(t) = \sum_j N_j(t)\phi_j \quad (j = n, n+1, \ldots) \tag{8.139}$$

within the time interval $(t_n, t_n + \Delta t_n)$. In the linear case with $j = n, n+1$ the interpolation functions are

$$N_n = 1 - \tau, \quad N_{n+1} = \tau \tag{8.140}$$

where

$$\tau = \frac{t - t_n}{\Delta t_n} \tag{8.141}$$

Inserting (8.139) into (8.129) a temporal weighted residual approximation, similar to a spatial weak statement (8.36), can be written

$$\int_{t_n}^{t_n+\Delta t_n} w_i(t)\Big[\boldsymbol{O}\cdot(\dot{N}_n\phi_n + \dot{N}_{n+1}\phi_{n+1}) + \boldsymbol{K}\cdot(N_n\phi_n + N_{n+1}\phi_{n+1}) - \boldsymbol{F}\Big]dt = 0 \tag{8.142}$$

with

$$\dot{N}_n = -\frac{1}{\Delta t_n}, \quad \dot{N}_{n+1} = \frac{1}{\Delta t_n} \tag{8.143}$$

where $w_i(t)$ represents the weighting function set to be specified below. If we substitute (8.141) with $dt = \Delta t_n d\tau$ into (8.142) and divide by Δt_n, the following time marching recurrence formula results

$$\begin{aligned}&\Big(\boldsymbol{O}\int_0^1 w_i \frac{1}{\Delta t_n} d\tau + \boldsymbol{K}\int_0^1 w_i \tau d\tau\Big) \cdot \boldsymbol{\phi}_{n+1} - \\ &\Big(\boldsymbol{O}\int_0^1 w_i \frac{1}{\Delta t_n} d\tau - \boldsymbol{K}\int_0^1 w_i (1-\tau) d\tau\Big) \cdot \boldsymbol{\phi}_n - \int_0^1 w_i \boldsymbol{F} d\tau = 0\end{aligned} \tag{8.144}$$

Introducing a weighting coefficient θ defined as

$$\theta = \frac{\int_0^1 w_i \tau d\tau}{\int_0^1 w_i d\tau} \tag{8.145}$$

and assuming a linear variation of $\boldsymbol{F}(t) \approx N_n(t)\boldsymbol{F}_n + N_{n+1}(t)\boldsymbol{F}_{n+1}$ within the time interval where

$$\frac{\int_0^1 w_i \boldsymbol{F} d\tau}{\int_0^1 w_i d\tau} = \boldsymbol{F}_{n+1}\theta + \boldsymbol{F}_n(1-\theta) \tag{8.146}$$

the final form of the recurrence scheme (8.144) is given by

$$\Big(\frac{\boldsymbol{O}}{\Delta t_n} + \boldsymbol{K}\theta\Big) \cdot \boldsymbol{\phi}_{n+1} = \Big(\frac{\boldsymbol{O}}{\Delta t_n} - \boldsymbol{K}(1-\theta)\Big) \cdot \boldsymbol{\phi}_n + \big(\boldsymbol{F}_{n+1}\theta + \boldsymbol{F}_n(1-\theta)\big) \tag{8.147}$$

to solve $\boldsymbol{\phi}_{n+1}$ at the new time plane $n+1$ from the preceding solution $\boldsymbol{\phi}_n$ at the previous time plane n. Using a Galerkin weighting with $w_i = N_{n+1}$ the weighting coefficient (8.145) yields $\theta = 2/3$.

8.13.4 The θ–Family of Time Integration Methods

Introducing a more general weighting coefficient ($0 \le \theta \le 1$), we can write

$$\begin{aligned}\boldsymbol{\phi}(t_n + \theta\Delta t_n) &= \theta\boldsymbol{\phi}(t_n + \Delta t_n) + (1-\theta)\boldsymbol{\phi}(t_n) \\ \boldsymbol{F}(t_n + \theta\Delta t_n) &= \theta\boldsymbol{F}(t_n + \Delta t_n) + (1-\theta)\boldsymbol{F}(t_n) \\ \dot{\boldsymbol{\phi}}(t_n + \theta\Delta t_n) &= \theta\dot{\boldsymbol{\phi}}(t_n + \Delta t_n) + (1-\theta)\dot{\boldsymbol{\phi}}(t_n)\end{aligned} \tag{8.148}$$

Using the Taylor series expansion for $\phi_{n+1} = \phi(t_n + \Delta t_n)$ about t_n and $\phi_n = \phi(t_{n+1} - \Delta t_n)$ about t_{n+1}, respectively,

$$\begin{aligned} \phi_{n+1} &= \phi_n + \Delta t_n \dot{\phi}_n + \tfrac{\Delta t_n^2}{2} \ddot{\phi}_n + \tfrac{\Delta t_n^3}{6} \dddot{\phi}_n + \ldots \\ \phi_n &= \phi_{n+1} - \Delta t_n \dot{\phi}_{n+1} + \tfrac{\Delta t_n^2}{2} \ddot{\phi}_{n+1} - \tfrac{\Delta t_n^3}{6} \dddot{\phi}_{n+1} + \ldots \end{aligned} \tag{8.149}$$

we obtain a forward difference approximation, called *forward Euler* (FE)

$$\dot{\phi}_n = \frac{\phi_{n+1} - \phi_n}{\Delta t_n} - \mathcal{O}(\Delta t_n) - \mathcal{O}(\Delta t_n^2) \tag{8.150}$$

and a backward difference approximation, called *backward Euler* (BE)

$$\dot{\phi}_{n+1} = \frac{\phi_{n+1} - \phi_n}{\Delta t_n} + \mathcal{O}(\Delta t_n) - \mathcal{O}(\Delta t_n^2) \tag{8.151}$$

which are accurate to a first-order truncation error of $\mathcal{O}(\Delta t_n)$. Inserting (8.151) and (8.150) into (8.148) it results[9]

$$\dot{\phi}(t_n + \theta \Delta t_n) = \frac{\phi_{n+1} - \phi_n}{\Delta t_n} + \mathcal{O}\big((\theta - \tfrac{1}{2})\Delta t_n, \Delta t_n^2\big) \tag{8.152}$$

We recognize from (8.152) that the difference approximation is of second-order accuracy of $\mathcal{O}(\Delta t_n^2)$ if (and only if) $\theta = \frac{1}{2}$, for all other values of θ within $(0 \leq \theta \leq 1)$ the difference approximation is accurate to a first-order truncation error of $\mathcal{O}(\Delta t_n)$.

Common time stepping schemes result if choosing θ in an appropriate manner, viz.,

$$\begin{aligned} \theta &= 0 && \text{explicit scheme, } \mathcal{O}(\Delta t_n) \\ \theta &= \tfrac{1}{2} && \text{trapezoid rule (Crank-Nicolson scheme), } \mathcal{O}(\Delta t_n^2) \\ \theta &= \tfrac{2}{3} && \text{Galerkin scheme, } \mathcal{O}(\Delta t_n) \\ \theta &= 1 && \text{implicit scheme, } \mathcal{O}(\Delta t_n) \end{aligned} \tag{8.153}$$

[9] The 2nd-order derivatives are obtained by repeated application of 1st-order approximations:

$$\ddot{\phi}_{n+1} = \frac{\dot{\phi}_{n+1} - \dot{\phi}_n}{\Delta t_n} + \mathcal{O}(\Delta t_n), \quad \ddot{\phi}_n = \frac{\dot{\phi}_{n+1} - \dot{\phi}_n}{\Delta t_n} - \mathcal{O}(\Delta t_n)$$

so that

$$\ddot{\phi}_{n+1} = \ddot{\phi}_n + \mathcal{O}(\Delta t_n)$$

Inserting (8.148) with (8.152) into (8.129) we obtain the following algebraic system of equations

$$\left(\frac{\boldsymbol{O}}{\Delta t_n} + \boldsymbol{K}\theta\right) \cdot \boldsymbol{\phi}_{n+1} = \left(\frac{\boldsymbol{O}}{\Delta t_n} - \boldsymbol{K}(1-\theta)\right) \cdot \boldsymbol{\phi}_n + \left(\boldsymbol{F}_{n+1}\theta + \boldsymbol{F}_n(1-\theta)\right) \tag{8.154}$$

and accordingly for the normalized form (8.131)

$$\left(\frac{\boldsymbol{\delta}}{\Delta t_n} + \boldsymbol{\mu}\theta\right) \cdot \boldsymbol{\phi}_{n+1} = \left(\frac{\boldsymbol{\delta}}{\Delta t_n} - \boldsymbol{\mu}(1-\theta)\right) \cdot \boldsymbol{\phi}_n + \left(\boldsymbol{f}_{n+1}\theta + \boldsymbol{f}_n(1-\theta)\right) \tag{8.155}$$

to recursively solve $\boldsymbol{\phi}_{n+1}$ at the new time plane $n+1$ from the preceding solution $\boldsymbol{\phi}_n$ at the previous time plane n, starting from the IC (8.130) at $n = 0$.

8.13.5 Predictor-Corrector Methods

A powerful alternative to the two-stage $\theta-$implicit/explicit recurrence solution (8.154) is the predictor-corrector method which was originally developed by Gresho et al. [209, 211, 212], hereafter referred to as GLS. This time integration method 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. The predictor-corrector methods provide a rational mathematical basis for adaptively selecting the time step via error control. Such an adaptive time stepping is clearly superior to procedures based exclusively on empirical relations, e.g., a target-based or heuristic time stepping control as discussed in [124,141,582]. In the present analysis both 1st- and 2nd-order accurate variable step predictor-corrector schemes are of interest. The 1st-order accurate scheme refers to an explicit forward Euler (FE) formula as the predictor and the implicit backward Euler (BE) method as the corrector. It will hereafter be termed as the FE/BE predictor-corrector scheme. For the 2nd-order accurate method the explicit method is based on the Adams-Bashforth (AB) predictor, while the trapezoid rule (TR) is used as corrector with 2nd-order accuracy. It will be hereafter called as the AB/TR predictor-corrector scheme. The schemes are applied to the system of equations (8.129) or (8.131) written in a simplified form:

$$\dot{\boldsymbol{\phi}} = \boldsymbol{r}(\boldsymbol{\phi}) \quad \text{with} \quad \boldsymbol{r}(\boldsymbol{\phi}) = \boldsymbol{f} - \boldsymbol{\mu} \cdot \boldsymbol{\phi} \tag{8.156}$$

8.13.5.1 GLS 1st-Order Forward Euler (FE)/Backward Euler (BE) Scheme

Predictor Solution

The FE scheme applied to $\dot{\boldsymbol{\phi}} = \boldsymbol{r}(\boldsymbol{\phi})$ gives, cf. (8.149),

$$\boldsymbol{\phi}_{n+1}^p = \boldsymbol{\phi}_n + \Delta t_n \boldsymbol{r}(\boldsymbol{\phi}_n) = \boldsymbol{\phi}_n + \Delta t_n \dot{\boldsymbol{\phi}}_n \tag{8.157}$$

where the superscript p indicates the predictor values at the new time plane $n+1$. The predictor provides a tentative solution at $n+1$.

Corrector Solution

The BE corrector scheme applied to (8.156) is

$$\boldsymbol{\phi}_{n+1} = \boldsymbol{\phi}_n + \Delta t_n \boldsymbol{r}(\boldsymbol{\phi}_{n+1}) = \boldsymbol{\phi}_n + \Delta t_n \dot{\boldsymbol{\phi}}_{n+1} \tag{8.158}$$

whose inversion yields the 'acceleration' vector

$$\dot{\boldsymbol{\phi}}_{n+1} = \frac{\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}_n}{\Delta t_n} \tag{8.159}$$

to be used for preparing the next predictor step on the RHS of (8.157). The corrector (8.158) provides the actual solution at $n+1$, which is commonly depart from the predictor solution (8.157).

Local Truncation Error (LTE) Estimation

The LTE $\boldsymbol{d}_{n+1}$ is defined as the residual

$$\boldsymbol{d}_{n+1} = \boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}(t_{n+1}) \tag{8.160}$$

between the approximate solution $\boldsymbol{\phi}_{n+1}$ and the exact solution $\boldsymbol{\phi}(t_{n+1})$ at the new time plane $n+1$. Practically, we determine the exact solution via Taylor series analysis and assume that the exact solution is available at the beginning of the time step. We obtain for the FE formula

$$\begin{aligned}\boldsymbol{d}_{n+1}^p &= \boldsymbol{\phi}_{n+1}^p - \boldsymbol{\phi}(t_{n+1}) \\ &= \boldsymbol{\phi}_n + \Delta t_n \dot{\boldsymbol{\phi}}_n - \Big(\boldsymbol{\phi}_n + \Delta t_n \dot{\boldsymbol{\phi}}_n + \tfrac{\Delta t_n^2}{2} \ddot{\boldsymbol{\phi}}_n + \tfrac{\Delta t_n^3}{6} \dddot{\boldsymbol{\phi}}_n + \ldots\Big) \\ &= -\tfrac{\Delta t_n^2}{2} \ddot{\boldsymbol{\phi}}_n + \mathcal{O}(\Delta t_n^3)\end{aligned} \tag{8.161}$$

and for the BE scheme

$$\begin{aligned}
\boldsymbol{d}_{n+1} &= \phi_{n+1} - \phi(t_{n+1}) \\
&= \phi_n + \Delta t_n \dot{\phi}_{n+1} - \Big(\phi_n + \Delta t_n \dot{\phi}_{n+1} - \tfrac{\Delta t_n^2}{2} \ddot{\phi}_{n+1} + \tfrac{\Delta t_n^3}{6} \dddot{\phi}_{n+1} - \ldots\Big) \\
&= \tfrac{\Delta t_n^2}{2} \ddot{\phi}_{n+1} + \mathcal{O}(\Delta t_n^3) \\
&= \tfrac{\Delta t_n^2}{2} \ddot{\phi}_n + \mathcal{O}(\Delta t_n^3)
\end{aligned} \tag{8.162}$$

taking into account that the exact solution is available at t_n (and not at t_{n+1}) by definition. From (8.161) and (8.162) it directly follows

$$\boldsymbol{d}_{n+1} = \frac{1}{2}(\phi_{n+1} - \phi^p_{n+1}) + \mathcal{O}(\Delta t_n^3) \tag{8.163}$$

by using $\boldsymbol{d}_{n+1} = -\boldsymbol{d}^p_{n+1} + \mathcal{O}(\Delta t_n^3)$. It provides an estimate of the LTE in a single BE step, where ϕ_{n+1} and ϕ^p_{n+1} are available from the corrector and predictor solution, respectively, at time plane $n + 1$.

Time Step Selection

On this basis we can determine a useful formula for the acceptable size of the next time step as follows. From (8.162) we find

$$\frac{\|\boldsymbol{d}_{n+2}\|}{\|\boldsymbol{d}_{n+1}\|} = \left(\frac{\Delta t_{n+1}}{\Delta t_n}\right)^2 \frac{\|\ddot{\phi}_{n+1}\|}{\|\ddot{\phi}_n\|} \tag{8.164}$$

where $\boldsymbol{d}_{n+1}$ is available from (8.163). The idea is now to keep the expected LTE at the next time plane $n + 2$ equal to a pre-set (tolerable, target) error measure ϵ, i.e., $\|\boldsymbol{d}_{n+2}\| = \epsilon$. Since $\ddot{\phi}_{n+1} = \ddot{\phi}_n + \mathcal{O}(\Delta t_n)$ (see[9]), (8.164) permits an estimate for the (potential) next time step size. Neglecting higher-order terms, we finally obtain from (8.164)

$$\Delta t_{n+1} = \Delta t_n \left(\frac{\epsilon}{\|\boldsymbol{d}_{n+1}\|}\right)^{1/2} \tag{8.165}$$

In this manner, the potential size of the next time step can be determined by the error norm $\|\boldsymbol{d}_{n+1}\|$ estimated from the difference between the predicted and corrected solutions in (8.163). It can be used as a RMS error norm $\|\boldsymbol{d}_{n+1}\|_{\text{RMS}}$, cf. (8.28), or as a maximum error norm $\|\boldsymbol{d}_{n+1}\|_{L_\infty}$, cf. (8.29).

8.13.5.2 GLS 2nd-Order Adams-Bashforth (AB)/Trapezoid Rule (TR) Scheme

Predictor Solution

The 2nd-order AB formula applied to $\dot{\phi} = \boldsymbol{r}(\phi)$ is (e.g., see [211])

$$\phi_{n+1}^{p} = \phi_n + \frac{\Delta t_n}{2}\left[\left(2 + \frac{\Delta t_n}{\Delta t_{n-1}}\right)\dot{\phi}_n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\phi}_{n-1}\right] \tag{8.166}$$

where $\Delta t_n = t_{n+1} - t_n$ and $\Delta t_{n-1} = t_n - t_{n-1}$. It represents an explicit two-step method and requires two history vectors of acceleration at the current time plane $\dot{\phi}_n$ and at the previous time plane $\dot{\phi}_{n-1}$. Since $\dot{\phi}_{n-1}$ is additionally needed, the AB formula cannot be applied before the second step ($n = 1$). Accordingly, the prediction has to be started with the FE predictor (8.157) and error estimation therefore begins at the completion of the second step.

Corrector Solution

The corrector step applied to (8.156) is based on the 2nd-order accurate TR, which reads

$$\phi_{n+1} = \phi_n + \frac{\Delta t_n}{2}\left[\boldsymbol{r}(\phi_n) + \boldsymbol{r}(\phi_{n+1})\right] = \phi_n + \frac{\Delta t_n}{2}\left(\dot{\phi}_n + \dot{\phi}_{n+1}\right) \tag{8.167}$$

whose inversion yields the history vector of acceleration

$$\dot{\phi}_{n+1} = \frac{2}{\Delta t_n}(\phi_{n+1} - \phi_n) - \dot{\phi}_n \tag{8.168}$$

to be used for preparing the next predictor step on the RHS of (8.166), where $\dot{\phi}_n$ could be available from the previous application of the same equation. However, Bixler [48] has shown that the previous accelaration vector $\dot{\phi}_n$ used in (8.168) can produce an oscillatory instability in the AB predictor in cases as a steady state is approached. Under such conditions $\phi_{n+1} - \phi_n$ in (8.168) will go to zero, however, $\dot{\phi}_n$ may not because of the recursive dependence on previous estimates of the acceleration vector. Bixler [48] proposed an alternative to $\dot{\phi}_n$ in (8.168) by the following finite difference relation[10]:

[10] Truncated Taylor series expansions for ϕ_{n-1} and ϕ_{n+1} about t_n give:

$$\phi_{n-1} = \phi_n - \Delta t_{n-1}\dot{\phi}_n + \frac{\Delta t_{n-1}^2}{2}\ddot{\phi}_n, \quad \phi_{n+1} = \phi_n + \Delta t_n\dot{\phi}_n + \frac{\Delta t_n^2}{2}\ddot{\phi}_n$$

$$\dot{\phi}_n = \frac{\Delta t_{n-1}}{\Delta t_n + \Delta t_{n-1}}\left(\frac{\phi_{n+1} - \phi_n}{\Delta t_n}\right) + \frac{\Delta t_n}{\Delta t_n + \Delta t_{n-1}}\left(\frac{\phi_n - \phi_{n-1}}{\Delta t_{n-1}}\right) \tag{8.169}$$

which is also $\mathcal{O}(\Delta t_n^2)$. Inserting (8.169) into (8.168) the following formula is used to compute the acceleration vector for the next AB step (8.166), viz.,

$$\dot{\phi}_{n+1} = \left(2 - \frac{\Delta t_{n-1}}{\Delta t_n + \Delta t_{n-1}}\right)\left(\frac{\phi_{n+1} - \phi_n}{\Delta t_n}\right) - \left(\frac{\Delta t_n}{\Delta t_n + \Delta t_{n-1}}\right)\left(\frac{\phi_n - \phi_{n-1}}{\Delta t_{n-1}}\right) \tag{8.170}$$

Local Truncation Error (LTE) Estimation

In analogy to the FE/BE scheme in Sect. 8.13.5.1 the LTE is obtained for the AB predictor

$$\begin{aligned} \boldsymbol{d}_{n+1}^p &= \phi_{n+1}^p - \phi(t_{n+1}) \\ &= \phi_n + \tfrac{\Delta t_n}{2}\big[\big(2 + \tfrac{\Delta t_n}{\Delta t_{n-1}}\big)\dot{\phi}_n - \tfrac{\Delta t_n}{\Delta t_{n-1}}\dot{\phi}_{n-1}\big] - \phi(t_{n+1}) \\ &= \phi_n + \tfrac{\Delta t_n}{2}\big[\big(2 + \tfrac{\Delta t_n}{\Delta t_{n-1}}\big)\dot{\phi}_n - \tfrac{\Delta t_n}{\Delta t_{n-1}}\big(\dot{\phi}_n - \Delta t_{n-1}\ddot{\phi}_n + \tfrac{\Delta t_{n-1}^2}{2}\dddot{\phi}_n - \ldots\big)\big] \\ &\quad -\Big(\phi_n + \Delta t_n\dot{\phi}_n + \tfrac{\Delta t_n^2}{2}\ddot{\phi}_n + \tfrac{\Delta t_n^3}{6}\dddot{\phi}_n + \mathcal{O}(\Delta t_n^4)\Big) \\ &= -\tfrac{1}{12}\big(2 + 3\tfrac{\Delta t_{n-1}}{\Delta t_n}\big)\Delta t_n^3\dddot{\phi}_n + \mathcal{O}(\Delta t_n^4) \end{aligned} \tag{8.171}$$

where the exact solution is used at $t_{n-1} = t_n - \Delta t_{n-1}$ to invoke Taylor series. Similarly, the LTE for the TR corrector results in

$$\begin{aligned} \boldsymbol{d}_{n+1} &= \phi_{n+1} - \phi(t_{n+1}) \\ &= \phi_n + \tfrac{\Delta t_n}{2}\big(\dot{\phi}_n + \dot{\phi}_{n+1}\big) - \phi(t_{n+1}) \\ &= \phi_n + \tfrac{\Delta t_n}{2}\big[\big(\dot{\phi}_{n+1} - \Delta t_n\ddot{\phi}_{n+1} + \tfrac{\Delta t_n^2}{2}\dddot{\phi}_{n+1} - \ldots\big) + \dot{\phi}_{n+1}\big] \\ &\quad -\Big(\phi_n + \Delta t_n\dot{\phi}_{n+1} - \tfrac{\Delta t_n^2}{2}\ddot{\phi}_{n+1} + \tfrac{\Delta t_n^3}{6}\dddot{\phi}_{n+1} - \mathcal{O}(\Delta t_n^4)\Big) \\ &= \tfrac{\Delta t_n^3}{12}\dddot{\phi}_{n+1} + \mathcal{O}(\Delta t_n^4) \\ &= \tfrac{\Delta t_n^3}{12}\dddot{\phi}_n + \mathcal{O}(\Delta t_n^4) \end{aligned} \tag{8.172}$$

Using the first expression to write $\dot{\phi}_n = \frac{\phi_n - \phi_{n-1}}{\Delta t_{n-1}} + \frac{\Delta t_{n-1}}{2}\ddot{\phi}_n$ and inserting into the second formula with $\frac{\Delta t_n}{2}\ddot{\phi}_n = \frac{\phi_{n+1} - \phi_n}{\Delta t_n} - \dot{\phi}_n$, we obtain

$$\dot{\phi}_n = \frac{\phi_n - \phi_{n-1}}{\Delta t_{n-1}} + \frac{\Delta t_{n-1}}{\Delta t_n}\left(\frac{\phi_{n+1} - \phi_n}{\Delta t_n} - \dot{\phi}_n\right)$$

After some manipulations we finally find

$$\dot{\phi}_n = \frac{\Delta t_{n-1}}{\Delta t_n + \Delta t_{n-1}}\left(\frac{\phi_{n+1} - \phi_n}{\Delta t_n}\right) + \frac{\Delta t_n}{\Delta t_n + \Delta t_{n-1}}\left(\frac{\phi_n - \phi_{n-1}}{\Delta t_{n-1}}\right)$$

From (8.171) and (8.172) we can directly express the LTE of a single TR step as

$$\boldsymbol{d}_{n+1} = \frac{\phi_{n+1} - \phi^{p}_{n+1}}{3\left(1 + \frac{\Delta t_{n-1}}{\Delta t_n}\right)} + \mathcal{O}(\Delta t_n^4) \tag{8.173}$$

providing a function of the available predictor solution ϕ^{p}_{n+1} and corrector solution ϕ_{n+1} at the time plane $n+1$.

Time Step Selection

In analogy to Sect. 8.13.5.1 we can estimate the next time step size for the AB/TR scheme based on the requirement that an error norm for the next step should equal a pre-set tolerance measure $\epsilon = \|\boldsymbol{d}_{n+2}\|$. From (8.172) we find

$$\frac{\|\boldsymbol{d}_{n+2}\|}{\|\boldsymbol{d}_{n+1}\|} = \left(\frac{\Delta t_{n+1}}{\Delta t_n}\right)^3 \frac{\|\dddot{\phi}_{n+1}\|}{\|\dddot{\phi}_n\|} \tag{8.174}$$

where $\boldsymbol{d}_{n+1}$ is known from (8.173). Neglecting higher-order terms and since $\dddot{\phi}_{n+1} = \dddot{\phi}_n + \mathcal{O}(\Delta t_n)$, we finally obtain from (8.174) the following relation

$$\Delta t_{n+1} = \Delta t_n \left(\frac{\epsilon}{\|\boldsymbol{d}_{n+1}\|}\right)^{1/3} \tag{8.175}$$

which is used to compute the potential next time step size.

8.13.5.3 Major Solution Steps and Tactic of Time Step Control

In Table 8.7 we summarize the major solution steps of the 1st-order accurate FE/BE and 2nd-order accurate AB/TR predictor-corrector schemes. In step 0 the time marching procedures are initialized by computing the acceleration vector $\dot{\phi}_0$ based on the IC (8.130): $\phi(t_0) = \phi_0$. Furthermore, an initial time step size Δt_0 is chosen, which should be kept sufficiently small. The error tolerance ϵ is the only user-specified parameter to control the entire adaptive time marching process. It has significant effect on cost and accuracy. A too large value of ϵ possesses a poor error estimate and the AB/TR becomes prone to oscillate when large time steps are used. Too small an ϵ, however, will make the (albeit accurate) computations unacceptably expensive. In practice, it has been shown in many applications that a relative error per time step of $\epsilon = 10^{-3}$–10^{-4} is quite optimal with respect to accuracy and performance. Note that a decrease of ϵ by one power will approximately double the total number of time steps.

The computations per each time step consist of five major solution steps as listed in Table 8.7. At first, the predictor solution ϕ^p_{n+1} at the new time plane $n+1$ has to be computed by using the explicit 1st-order accurate FE and 2nd-order accurate AB schemes. The AB scheme must start at $n+1$ as the FE scheme because the required acceleration vector $\dot{\phi}_{n-1}$ is only available from the second step onward. Only at the second step $n = 2$ the usual AB predictor procedure is started. All the predictors are cheaply computable and their extra effort is small. While these predictors are subsequently needed to estimate the truncation error for the time step control, they are also useful to linearize the governing PDE in the presence of nonlinearities. Step 2, the corrector ϕ_{n+1}, is the actual solution of the governing PDE, $\boldsymbol{O} \cdot \dot{\phi} + \boldsymbol{K} \cdot \phi = \boldsymbol{F}$, via the implicit BE and TR schemes. If we additionally admit nonlinear dependencies in the form

$$\boldsymbol{O}(\phi) \cdot \dot{\phi} + \boldsymbol{K}(\phi) \cdot \phi = \boldsymbol{F}(\phi) \tag{8.176}$$

we can solve the linear system by using the predictor solution ϕ^p_{n+1} at the new time plane $n+1$

$$\Big(\frac{\boldsymbol{O}(\phi^p_{n+1})}{\theta \Delta t_n} + \boldsymbol{K}(\phi^p_{n+1})\Big) \cdot \phi_{n+1} = \boldsymbol{O}(\phi^p_{n+1}) \cdot \Big[\frac{1}{\theta \Delta t_n}\phi_n + \big(\tfrac{1}{\theta} - 1\big)\dot{\phi}_n\Big] + \boldsymbol{F}_{n+1}(\phi^p_{n+1}) \tag{8.177}$$

where $\theta = \frac{1}{2}$ for the TR scheme and $\theta = 1$ for the BE scheme.

Once the corrector solution ϕ_{n+1} is available at the new time plane $n+1$, in step 3 the acceleration vectors $\dot{\phi}_{n+1}$ can be computed, which will be needed in the following $n+2$ time step in the predictor and AB-corrector. With the known predictor and corrector solutions, ϕ^p_{n+1} and ϕ_{n+1}, respectively, in step 4 the LTE $\boldsymbol{d}_{n+1}$ is determined at the current time plane $n+1$. Appropriate error norms are applied to the vector $\boldsymbol{d}_{n+1}$. Commonly, the RMS L_2 error norm (8.28)

$$\|\boldsymbol{d}_{n+1}\|_{\mathrm{RMS}} = \Bigg[\frac{1}{N_{\mathrm{P}}}\bigg(\frac{1}{\phi^2_{\max}}\sum_{j=1}^{N_{\mathrm{P}}} d^2_{j,n+1}\bigg)\Bigg]^{\frac{1}{2}} \tag{8.178}$$

or the maximum L_∞ error norm (8.29)

$$\|\boldsymbol{d}_{n+1}\|_{L_\infty} = \frac{1}{\phi_{\max}} \max_j |d_{j,n+1}| \tag{8.179}$$

are chosen, where $\phi_{\max}$ is the maximum value of the state variable ϕ_{n+1} detected at the time plane $n+1$ to normalize the error vectors.

In step 5, the potential next time step size Δt_{n+1} is determined by using the just estimated error norm $\|\boldsymbol{d}_{n+1}\| \in (\|\boldsymbol{d}_{n+1}\|_{\mathrm{RMS}}, \|\boldsymbol{d}_{n+1}\|_{L_\infty})$ with the user-

Table 8.7 Summarized solution steps of the 1st-order accurate FE/BE and 2nd-order accurate AB/TR predictor-corrector schemes applied to $\boldsymbol{O}\cdot\dot{\phi}+\boldsymbol{K}\cdot\phi = \boldsymbol{F}$, (8.129)

	Step	n	FE/BE scheme $\mathcal{O}(\Delta t_n)$	AB/TR scheme $\mathcal{O}(\Delta t_n^2)$
0	Initialization	$n = 0$	$\boldsymbol{O}\cdot\dot{\phi}_0 = \boldsymbol{F} - \boldsymbol{K}\cdot\phi_0$ with given $\phi_0, \Delta t_0, \epsilon$	$\boldsymbol{O}\cdot\dot{\phi}_0 = \boldsymbol{F} - \boldsymbol{K}\cdot\phi_0$ with given $\phi_0, \Delta t_0, \epsilon$
1	Predictor	$n + 1$	$\phi^p_{n+1} = \phi_n + \Delta t_n\dot{\phi}_n$	$\phi^p_{n+1} = \begin{cases} \phi_n + \frac{\Delta t_n}{2}\big[\big(2 + \frac{\Delta t_n}{\Delta t_{n-1}}\big)\dot{\phi}_n \\ -\frac{\Delta t_n}{\Delta t_{n-1}}\dot{\phi}_{n-1}\big] & \text{for} \quad n > 1 \\ \phi_n + \Delta t_n\dot{\phi}_n & \text{for} \quad n = 1 \end{cases}$
2	Corrector		$\big(\frac{\boldsymbol{O}}{\Delta t_n} + \boldsymbol{K}\big)\cdot\phi_{n+1} =$ $\frac{\boldsymbol{O}}{\Delta t_n}\cdot\phi_n + \boldsymbol{F}_{n+1}$	$\big(\frac{2\boldsymbol{O}}{\Delta t_n} + \boldsymbol{K}\big)\cdot\phi_{n+1} =$ $\boldsymbol{O}\cdot\big(\frac{2}{\Delta t_n}\phi_n + \dot{\phi}_n\big) + \boldsymbol{F}_{n+1}$
3	Updated accelerations		$\dot{\phi}_{n+1} = \dfrac{\phi_{n+1} - \phi_n}{\Delta t_n}$	$\dot{\phi}_{n+1} = \big(2 - \dfrac{\Delta t_{n-1}}{\Delta t_n + \Delta t_{n-1}}\big)\big(\dfrac{\phi_{n+1} - \phi_n}{\Delta t_n}\big)$ $-\big(\frac{\Delta t_n}{\Delta t_n + \Delta t_{n-1}}\big)\big(\frac{\phi_n - \phi_{n-1}}{\Delta t_{n-1}}\big)$
4	Error estimation		$\boldsymbol{d}_{n+1} = \frac{1}{2}(\phi_{n+1} - \phi^p_{n+1})$	$\boldsymbol{d}_{n+1} = \dfrac{\phi_{n+1} - \phi^p_{n+1}}{3\big(1 + \frac{\Delta t_{n-1}}{\Delta t_n}\big)}$
5	Time step control		$\Delta t_{n+1} = \Delta t_n\left(\dfrac{\epsilon}{\Vert\boldsymbol{d}_{n+1}\Vert}\right)^{1/2}$	$\Delta t_{n+1} = \Delta t_n\left(\dfrac{\epsilon}{\Vert\boldsymbol{d}_{n+1}\Vert}\right)^{1/3}$
			if $\quad \Delta t_{n+1} \geq \Delta t_n$	solution ϕ_{n+1} accepted, proceed with increased time step at $n + 2$,
			else if $\quad \gamma\Delta t_n \leq \Delta t_{n+1} < \Delta t_n$ $(\gamma = 0.85)$	solution ϕ_{n+1} accepted, but proceed with unchanged time step $\Delta t_{n+1} = \Delta t_n$ at $n + 2$,
			else if $\quad \Delta t_{n+1} < \gamma\Delta t_n$	solution ϕ_{n+1} cannot be accepted and must be rejected, repeat at $n + 1$ with reduced time step $\Delta t_n = \Delta t^{\text{red}}_{n+1}$, where $\Delta t^{\text{red}}_{n+1} = \frac{\Delta t_n^2}{\Delta t_{n+1}}\left(\frac{\epsilon}{\Vert\boldsymbol{d}^{n+1}\Vert}\right)^{\varsigma}$ ($\varsigma = 1$ for FE/BE and $\varsigma = 2/3$ for AB/TR)

supplied error tolerance ϵ. The following criteria are used to monitor the progress of solution:

1. If

$$\varDelta t_{n+1} \geq \varDelta t_n \tag{8.180}$$

the current solution ϕ_{n+1} is accurate within the error bound defined by ϵ and the increase of the time step is always accepted. In practice it has shown to be useful that the increase of the time step should be optionally constrained by further conditions. Firstly, the time step should not exceed a prescribed maximum size, i.e., $\varDelta t_{n+1} \leq \varDelta t_{\max}$. Secondly, the rate for changing the time step size $\varXi = \varDelta t_{n+1}/\varDelta t_n$ has also to be limited, where $\varXi > 1$ can be 2, 3 or even more. Those constraints are beneficial to prevent inefficient oscillations in the time step size prediction. Then, the actually increased new time step is determined from

$$\varDelta t_{n+1}^{\text{actual}} = \min(\varDelta t_{n+1}, \varDelta t_{\max}, \varXi \varDelta t_n) \tag{8.181}$$

provided that $\varDelta t_{n+1}^{\text{actual}} \geq \varDelta t_n$.

2. Else if

$$\gamma \varDelta t_n \leq \varDelta t_{n+1} < \varDelta t_n \tag{8.182}$$

where γ is typically 0.85, the solution ϕ_{n+1} is accepted but the time step size is not changed, i.e., $\varDelta t_{n+1} = \varDelta t_n$.

3. Else if

$$\varDelta t_{n+1} < \gamma \varDelta t_n \tag{8.183}$$

the solution ϕ_{n+1} cannot be accepted within the required error tolerance ϵ and has to be rejected. The current time step must be repeated with a reduced time step size. The reduced time step is computed from (8.165) and (8.175), respectively, by replacing $\|\boldsymbol{d}\|_{n+1}$ and $\varDelta t_n$ with the just estimated $\|\boldsymbol{d}\|_{n+2}$ and $\varDelta t_{n+1}$ to obtain

$$\varDelta t_{n+1}^{\text{red}} = \frac{\varDelta t_n^2}{\varDelta t_{n+1}} \Big(\frac{\epsilon}{\|\boldsymbol{d}^{n+1}\|} \Big)^{\varsigma} \tag{8.184}$$

where $\varsigma = 1$ for FE/BE and $\varsigma = 2/3$ for AB/TR scheme. The new solution restarted with this smaller time step is again tested against the error conditions and further step reduction can follow. However, up to 12 such reduction cycles are only allowed, then the algorithm signals to restart the overall time stepping procedure under stronger error bounds and initial time step (e.g., decrease ϵ and/or $\varDelta t_0$).

After finishing solution step 5 for an acceptable solution ϕ_{n+1}, the time stepping procedure proceeds to the next time plane $n+2$, where it begins again with step 1 of Table 8.7. With the proposed predictor-corrector technique we can vary the size of the time step based solely on temporal accuracy requirements. Such an error-controlled adaptive time step selection strategy can follow the 'physics' of the underlying processes more intelligently and efficiently in comparison to heuristic rules. For example, the physics may require a small time step to follow a steep concentration or temperature profile over certain times or to adapt a sudden change in transient BC's, while at later times it may be sufficient to follow a slow development of a flow or transport regime in time with reasonably large time steps. In either case, the predictor-corrector algorithm will usually automatically select the appropriate time step in a reliable manner, where the time step is increased whenever possible and decreased only when necessary.

8.13.6 Stability Properties

Any of the time marching recurrence schemes derived above can be written for the homogeneous solution (i.e., the source/sink of error is unimportant in the context so that we can assume $\boldsymbol{F} = \boldsymbol{0}$) in the form

$$\boldsymbol{\phi}_{n+1} = \boldsymbol{A} \cdot \boldsymbol{\phi}_n \tag{8.185}$$

where $\boldsymbol{A}$ is the *amplification matrix*, which is given for the exact solution by the exponential decay relation (8.136)

$$\boldsymbol{A} = e^{-\boldsymbol{\mu}\Delta t_n}, \quad \boldsymbol{\mu} = \boldsymbol{O}^{-1} \cdot \boldsymbol{K} \tag{8.186}$$

and for the introduced time stepping schemes by the approximation

$$\begin{aligned} \boldsymbol{A} &= \left[\boldsymbol{O} + \theta \boldsymbol{K} \Delta t_n\right]^{-1} \cdot \left[\boldsymbol{O} - (1-\theta)\boldsymbol{K}\Delta t_n\right] \\ &= \left[\boldsymbol{\delta} + \theta \boldsymbol{\mu} \Delta t_n\right]^{-1} \cdot \left[\boldsymbol{\delta} - (1-\theta)\boldsymbol{\mu}\Delta t_n\right] \end{aligned} \tag{8.187}$$

in which θ identifies the different recurrence algorithms. Table 8.5 lists the preferred linear single-step operators for particular θ values. It is obvious as $\boldsymbol{A}$ is recursively applied to each new vector $\boldsymbol{\phi}_n$, the stability of the time integration method requires that any occurring approximation error must ultimately decay. Thus, $\boldsymbol{A}$ must be a *bounded* operator and the time integration scheme is considered *stable* for $|\boldsymbol{A}| < 1$.

The stability of the time integration approximations can be further analyzed via modal decomposition. The solution $\boldsymbol{\phi}$ is expressed in terms of its linearly independent eigenvectors and eigenvalues by

$$\boldsymbol{\phi} = \sum_{i=1}^{N_\mathrm{P}} \boldsymbol{\varphi}_i e^{-\boldsymbol{\delta}\lambda_i t} \tag{8.188}$$

where $\boldsymbol{\varphi}_i$ are the eigenvectors and λ_i are the eigenvalues. Applying (8.188) to $\dot{\boldsymbol{\phi}} + \boldsymbol{\mu} \cdot \boldsymbol{\phi} = \mathbf{0}$, (8.131), with $\boldsymbol{f} = \mathbf{0}$, it leads to the eigenproblem in the form

$$(\boldsymbol{\mu} - \boldsymbol{\delta}\lambda_i) \cdot \boldsymbol{\varphi}_i = \mathbf{0}, \qquad \forall i. \tag{8.189}$$

Since the eigenvectors have the properties of *modal orthogonality* in the form $\boldsymbol{\varphi}_j^T \cdot (\boldsymbol{\delta} \cdot \boldsymbol{\varphi}_i) = \delta_{ij}$, we find after multiplying (8.189) by $\boldsymbol{\varphi}_j^T$

$$\boldsymbol{\varphi}_j^T \cdot (\boldsymbol{\mu} \cdot \boldsymbol{\varphi}_i) = \lambda_i \delta_{ij} \tag{8.190}$$

showing that the eigenvectors are also orthogonal with respect to $\boldsymbol{\mu}$. Now, we assume that the semidiscrete solution can be approximated in terms of the eigenvectors as

$$\boldsymbol{\phi} = \sum_{i=1}^{N_\mathrm{P}} \boldsymbol{\varphi}_i y_i(t) \tag{8.191}$$

where $y_i(t)$ represent the mode participation factors to be determined. Substituting (8.191) into $\dot{\boldsymbol{\phi}} + \boldsymbol{\mu} \cdot \boldsymbol{\phi} = \mathbf{0}$, premultiplying with $\boldsymbol{\varphi}_j^T$ and applying the modal orthogonality conditions, leads to the result

$$\begin{aligned} \dot{y}_i \left[\boldsymbol{\varphi}_j^T \cdot (\boldsymbol{\delta} \cdot \boldsymbol{\varphi}_i)\right] + y_i \left[\boldsymbol{\varphi}_j^T \cdot (\boldsymbol{\mu} \cdot \boldsymbol{\varphi}_i)\right] &= 0 \quad \text{or} \\ \dot{y}_i \delta_{ij} + y_i \lambda_i \delta_{ij} &= 0 \quad \text{or} \\ \dot{y}_i + y_i \lambda_i &= 0 \quad \forall i. \end{aligned} \tag{8.192}$$

This modal formulation is very advantageous because it decouples the original equation into a sequence of scalar evolution equations for each mode $i = 1, \ldots, N_\mathrm{P}$. By applying the above time integration techniques, same as used for the original problem, now to the modal equations (8.192) we obtain similar to (8.185)

$$(y_i)_{n+1} = A_i (y_i)_n \tag{8.193}$$

where A_i is the scalar *amplification factor* of the ith mode, which is given for the exact solution by

$$A_i = e^{-\lambda_i \Delta t_n} \tag{8.194}$$

and for the introduced time stepping schemes by

$$A_i = \frac{1 - (1 - \theta)\lambda_i \Delta t_n}{1 + \theta \lambda_i \Delta t_n} \tag{8.195}$$

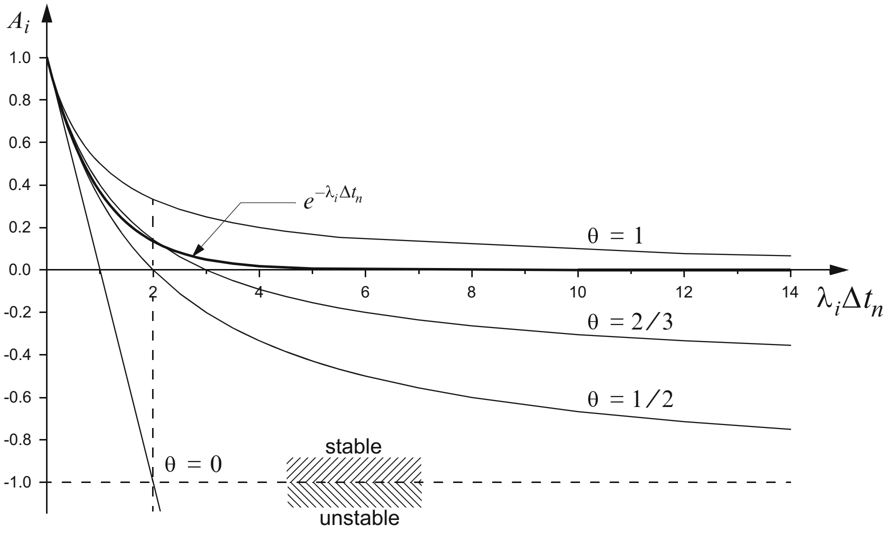


Fig. 8.19 Amplification factor A_i (decay function) for various linear θ−approximants (8.195) in comparison to the exact solution $e^{-\lambda_i \Delta t_n}$ of a mode i with eigenvalue λ_i

providing a scalar analog to (8.186) and (8.187), respectively. Taking into account that the eigenvalues λ_i cover the full eigenspectrum ranging between a maximum eigenvalue $\lambda_{\max}$ and a minimum eigenvalue $\lambda_{\min}$

$$\lambda_i = (\lambda_{\max}, \ldots, \lambda_{\min}) \tag{8.196}$$

which can comprise several orders of magnitude, the requirement for stability is that the amplification factor A_i must be $|A_i| < 1$, i.e.,

$$-1 < \frac{1 - (1-\theta)\lambda_i \Delta t_n}{1 + \theta \lambda_i \Delta t_n} < 1 \tag{8.197}$$

holding for all eigenvalues λ_i of the system.

Figure 8.19 illustrates how the amplification factor A_i of a mode i varies with $\lambda_i \Delta t_n$ for various θ of the four difference operators in comparison to the exact exponential decay $e^{\lambda_i \Delta t_n}$, where $\theta \in (0, \frac{1}{2}, \frac{2}{3}, 1)$ represents the explicit FE, the implicit TR (Crank-Nicolson), the implicit Galerkin and the fully implicit BE scheme, respectively. We easily recognize that the right-hand inequality of (8.197) imposes no restrictions on values of $\lambda_i \Delta t_n$ or θ. However, the left-hand inequality requires for stability that

$$(1 - 2\theta)\lambda_i \Delta t_n < 2 \tag{8.198}$$

when $\theta < \frac{1}{2}$. On the other hand, we see that there are no restrictions for $\theta \geq \frac{1}{2}$. Such algorithms satisfying (8.198) independent of the chosen time step size Δt_n are called *conditionally stable* ($\mathcal{A}-$stable). The TR (Crank-Nicolson) ($\theta = \frac{1}{2}$), the Galerkin scheme ($\theta = \frac{2}{3}$) and the implicit BE method ($\theta = 1$) belong to this category. In contrast, the explicit FE scheme ($\theta = 0$) is stable only if $\lambda_i \Delta t_n < 2$, otherwise A_i predicts unbounded (unstable) solutions with $A_i \to -\infty$ as evidenced in Fig. 8.19. Therefore, an explicit scheme is not an $\mathcal{A}-$stable method. As also shown in Fig. 8.19 the 2nd-order accurate TR scheme ($\theta = \frac{1}{2}$) fits very well with the exact solution and furnishes highest accuracy for $\lambda_i \Delta t_n < 1$ in comparison to all other linear single-step schemes. The implicit BE scheme ($\theta = 1$) approaches to the exact solution $A_i \to 0$ for very large time steps $\lambda_i \Delta t_n \to \infty$, while the Galerkin and the TR schemes satisfy $A_i \to -1$. Apparently, the Galerkin method ($\theta = \frac{2}{3}$) exhibits an optimal approximation behavior over the entire $\lambda_i \Delta t_n-$range.

The $\mathcal{A}-$stable time stepping algorithms satisfying (8.198) ensure boundedness and thus unconditional stability independently of the time step Δt_n. However, $\mathcal{A}-$stability is not sufficient to ensure smooth and wiggle-free (nonoscillatory) solutions. In fact, all algorithms which admit a negative amplification A_i (see Fig. 8.19) are prone to oscillatory behaviors if Δt_n becomes too large. The condition for nonoscillation (called $\mathcal{L}-$stability) requires $0 < A_i < 1$. The bound $A_i > 0$ gives with (8.195) the criterion

$$(1-\theta)\lambda_i \Delta t_n < 1 \tag{8.199}$$

to ensure nonoscillatory solutions. It is obvious that only the BE scheme ($\theta = 1$) can satisfy this condition for arbitrary step sizes Δt_n assuring $A_i \to 0$ for $\lambda_i \Delta t_n \to \infty$. Unlikely, in the TR scheme ($\theta = \frac{1}{2}$) the time step has to be restricted by a critical time step Δt_n^{crit} such as

$$\Delta t_n < \Delta t_n^{\text{crit}} = \frac{2}{\lambda_{\max}} \tag{8.200}$$

to ovoid oscillations in the solution (known as Crank-Nicolson noise [568]), which must be controlled by the maximum eigenvalue $\lambda_{\max}$.

Different methods exist for analyzing stability. One is the matrix method in which the eigenvalues of the matrix are estimated. To get a first (but simple) assessment of characteristic eigenvalues λ_i which are important to determine time step limitations, such as (8.200), we can use $\boldsymbol{\delta\lambda}_i = \boldsymbol{\mu}$ from (8.189) and estimate $\boldsymbol{\mu} = \boldsymbol{O}^{-1} \cdot \boldsymbol{K}$ on an element level basis, e.g., [590]. Let us consider for the sake of simplicity the 1D linear element (cf. Table G.1a of Appendix G), for which the element matrices have been derived in Appendix H. We find for the diagonal contributions of the (lumped) mass matrix O_{ii} and diffusion matrix C_{ii} from (H.7) at a (global) mode i, assuming a uniform meshing with element length Δx and constant parameters (storage – $\acute{\mathcal{R}}$, diffusion – D):

$$\lambda_i \approx \frac{C_{ii}}{O_{ii}} = \frac{2D}{\Delta x} \Big/ \frac{\acute{\mathcal{R}}\Delta x}{2}\Big|_{\mathrm{LM}} = \frac{4D}{\acute{\mathcal{R}}\Delta x^2} \tag{8.201}$$

written without advection, where we also drop the source/sink terms appearing in (H.7) because stability is independent of the forcing functions. Then, the assessment of condition (8.200) for the TR scheme yields[11]

$$\Delta t_n < \Delta t_n^{\mathrm{crit}} \approx \frac{\acute{\mathcal{R}}\Delta x^2}{2D} \tag{8.202}$$

which indicates that the critical time step is proportional to Δx^2 for a diffusion-dominant problem $D > 0$. We see that the smallest element size Δx dictates the criterion. In practice, however, the TR (or Crank-Nicolson) criterion (8.202) is commonly insignificant. Possible oscillations produced by the TR scheme are strictly bounded and small if linear finite elements are used. This is exemplified in Fig. 8.20 illustrating slight, but quickly damped oscillations in the temporal development of the solution for the TR (Crank-Nicolson) scheme using a constant time step larger than the critical step size (8.202). It is shown in [568] that Crank-Nicolson noise is more significant for finite elements of quadratic or higher-order type.

For a further analysis let us consider the spatio-temporal discretization of the system (8.154) for a simplified 1D problem. Again, we use linear elements thoroughly described in Sect. H.1 of Appendix H for a 1D domain as shown in Fig. 8.21. As indicated in Fig. 8.21 the assembly of the elements leads to a tridiagonal global matrix, where the final discrete equations can be expressed for the row of global interior node i by using the specific matrix entries of (H.8) and (H.10). For simplicity we drop source/sink and boundary terms and find for (8.154) the following discrete equations of the 3-node $(i+1, i, i-1)$−stencil:

$$\begin{gathered}\Big[\tfrac{\acute{\mathcal{R}}}{6} - \Delta t_n\theta\Big(\tfrac{D}{\Delta x^2} - \tfrac{q}{2\Delta x} - \tfrac{\vartheta}{6}\Big)\Big]\phi_{i+1,n+1} + \Big[\tfrac{2\acute{\mathcal{R}}}{3} + \Delta t_n\theta\Big(\tfrac{2D}{\Delta x^2} + \tfrac{2\vartheta}{3}\Big)\Big]\phi_{i,n+1} + \\ \Big[\tfrac{\acute{\mathcal{R}}}{6} - \Delta t_n\theta\Big(\tfrac{D}{\Delta x^2} + \tfrac{q}{2\Delta x} - \tfrac{\vartheta}{6}\Big)\Big]\phi_{i-1,n+1} = \Big[\tfrac{\acute{\mathcal{R}}}{6} + \Delta t_n(1-\theta)\Big(\tfrac{D}{\Delta x^2} - \tfrac{q}{2\Delta x} - \tfrac{\vartheta}{6}\Big)\Big]\phi_{i+1,n} + \\ \Big[\tfrac{2\acute{\mathcal{R}}}{3} - \Delta t_n(1-\theta)\Big(\tfrac{2D}{\Delta x^2} + \tfrac{2\vartheta}{3}\Big)\Big]\phi_{i,n} + \Big[\tfrac{\acute{\mathcal{R}}}{6} + \Delta t_n(1-\theta)\Big(\tfrac{D}{\Delta x^2} + \tfrac{q}{2\Delta x} - \tfrac{\vartheta}{6}\Big)\Big]\phi_{i-1,n}\end{gathered} \tag{8.203}$$

written for the CM matrix and

[11] Since $O_{ii}^e = \frac{\acute{\mathcal{R}}\Delta x}{3}\big|_{\mathrm{CM}}$ for a consistent mass (CM) matrix, the critical time step becomes even smaller:

$$\Delta t_n < \Delta t_n^{\mathrm{crit}} \approx \frac{\acute{\mathcal{R}}\Delta x^2}{3D}$$

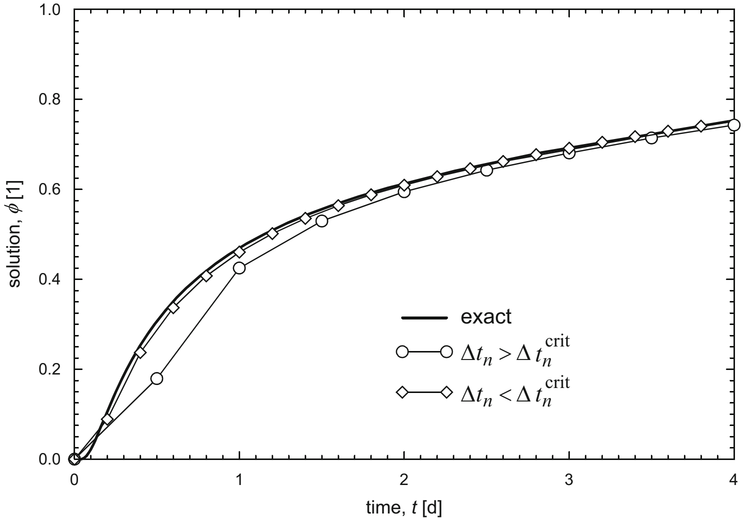


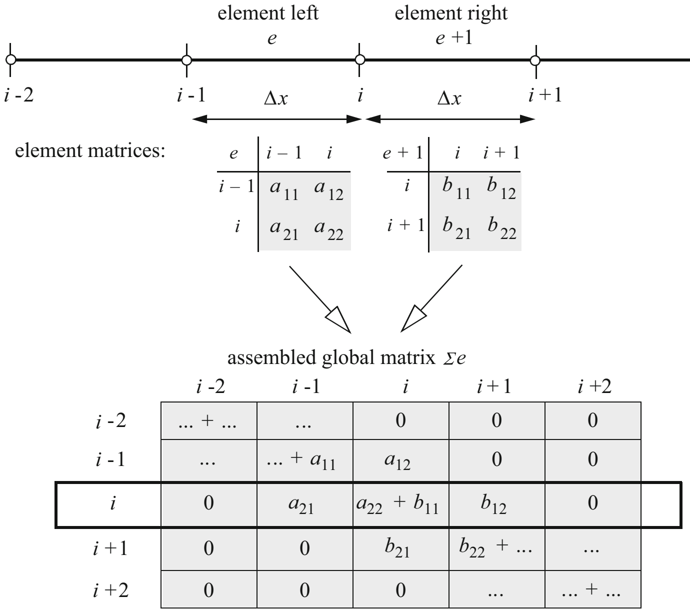
Fig. 8.20 Example problem of a transient diffusion in a 1D domain of $L = 1$ m discretized by five linear elements with $\Delta x = 0.2$ m, $\acute{\mathcal{R}} = 1$ and $D = 10^{-6}$ m^2 s^{-1} showing the history of Crank-Nicolson approximate solution $\phi(t)$ at $x = 0.3$ m in time for a constant time step Δt_n larger and smaller than the critical time step $\Delta t_n^{\text{crit}} = 0.23$ d, (8.202), in comparison to the exact solution

$$-\Delta t_n \theta\Big(\frac{D}{\Delta x^2} - \frac{q}{2\Delta x} - \frac{\vartheta}{6}\Big)\phi_{i+1,n+1} + \Big[\acute{\mathcal{R}} + \Delta t_n \theta\Big(\frac{2D}{\Delta x^2} + \frac{2\vartheta}{3}\Big)\Big]\phi_{i,n+1} - \Delta t_n \theta\Big(\frac{D}{\Delta x^2} + \frac{q}{2\Delta x} - \frac{\vartheta}{6}\Big)\phi_{i-1,n+1} = \Delta t_n(1-\theta)\Big(\frac{D}{\Delta x^2} - \frac{q}{2\Delta x} - \frac{\vartheta}{6}\Big)\phi_{i+1,n} + \Big[\acute{\mathcal{R}} - \Delta t_n(1-\theta)\Big(\frac{2D}{\Delta x^2} + \frac{2\vartheta}{3}\Big)\Big]\phi_{i,n} + \Delta t_n(1-\theta)\Big(\frac{D}{\Delta x^2} + \frac{q}{2\Delta x} - \frac{\vartheta}{6}\Big)\phi_{i-1,n} \tag{8.204}$$

written for the LM matrix, where the diffusion D, the advective flux q, the storage $\acute{\mathcal{R}}$, the decay rate ϑ and the length of the linear 1D element Δx are assumed constant.

Based on discrete equations such as in form of (8.203) or (8.204) a very common and most useful method for analyzing stability is the classical Fourier analysis, called *von Neumann stability analysis*, e.g., [149, 209, 376]. Von Neumann stability results necessary conditions at least on a uniform mesh, regardless of BC's. On this basis it can be shown that nonoscillatory solutions for the TR (or Crank-Nicolson) method are bound to the pair of inequalities

$$Cr < Pg < 1/Cr, \quad \text{or} \quad Cr < Pg \quad \text{and} \quad Cr < 1/Pg \tag{8.205}$$



e	$i-1$	i
$i-1$	a_{11}	a_{12}
i	a_{21}	a_{22}

$e+1$	i	$i+1$
i	b_{11}	b_{12}
$i+1$	b_{21}	b_{22}

	i -2	i -1	i	i +1	i +2
i -2	... + ...	...	0	0	0
i -1	...	... + a_{11}	a_{12}	0	0
i	0	a_{21}	a_{22} + b_{11}	b_{12}	0
i +1	0	0	b_{21}	b_{22} + ...	...
i +2	0	0	0	...	... + ...

Fig. 8.21 Node numbering and assembly to a tridiagonal global matrix for a uniform mesh of 1D linear elements

in which

$$Cr = \frac{q^* \Delta t_n}{\Delta x}, \quad \text{with} \quad q^* = q/\acute{\mathcal{R}} \tag{8.206}$$

and

$$Pg = \frac{q^* \Delta x}{2D^*}, \quad \text{with} \quad D^* = D/\acute{\mathcal{R}} \tag{8.207}$$

derived for the LM matrix, where Cr is the *Courant number* (named after the famous paper by Courant et al. [105]) and Pg defines the grid (mesh) *Péclet number*. The first limit $Cr < Pg$ is the 'diffusion limit' $\Delta t_n < \acute{\mathcal{R}}\Delta x^2/(2D)$ when $Pg < 1$ as already stated in (8.202) and the second one $PgCr < 1$ represents the 'advection-diffusion limit' $\Delta t_n < 2\acute{\mathcal{R}}D/q^2$ when $Pg \geq 1$. The second Crank-Nicolson criterion $PgCr < 1$ was also found by Perrochet and Bérod [415] by using a matrix method.

While the discrete equations (8.203) and (8.204) are $\mathcal{A}-$stable for $\theta \geq \frac{1}{2}$, i.e., stability is guaranteed for any time step Δt_n, nonoscillatory results require additional limits which directly follow from (8.203) and (8.204). It can be easily seen from (8.203) that the term $[\frac{\acute{\mathcal{R}}}{6} - \Delta t_n \theta(\frac{D}{\Delta x^2} + \frac{|q|}{2\Delta x} - \frac{\vartheta}{6})]$ must be negative to avoid oscillations in a CM formulation. It leads to a restriction for a minimum time step size, viz.,

$$\Delta t_n > \frac{\acute{\mathcal{R}} \Delta x^2}{\theta(6D + 3|q|\Delta x - \vartheta \Delta x^2)} \quad \text{for} \quad \theta \geq \tfrac{1}{2} \tag{8.208}$$

and additionally it should be required that $(6D + 3|q|\Delta x - \vartheta \Delta x^2) > 0$ which arises a further constraint related to the decay rate

$$\vartheta < \frac{3}{\Delta x^2}(2D + |q|\Delta x) \tag{8.209}$$

or a limit for the element length determined by the decay rate ϑ

$$\Delta x < \frac{3|q| + \sqrt{9q^2 + 24D\vartheta}}{2\vartheta} \quad \text{if} \quad \vartheta > 0 \tag{8.210}$$

The physical interpretation of a minimum time step size (8.208) for a consistent mass formulation is that the mesh is too coarse to transmit and distribute a quantity to the nodes of an element in a very short time interval so that bounded oscillations become unavoidable. However, we observe from (8.204) that such a minimum time step constraint does not exist for mass lumping, since $[-\Delta t_n \theta(\frac{D}{\Delta x^2} + \frac{|q|}{2\Delta x} - \frac{\vartheta}{6})]$ is always negative here, provided that (8.209) or (8.210) are satisfied. That means, the restriction for the decay rate (8.209) or (8.210) are present both for CM and LM matrix formulations. [Note that it could be possible to lump also the reaction matrix term similar to the LM matrix as discussed in Sect. 8.13.2, then the restrictions (8.209) or (8.210) would disappear.] We illustrate in Fig. 8.22 for a 1D example problem the oscillatory effect of CM if the time step is too small violating (8.208) and that mass lumping produces nonoscillatory results for the same time step. In practice, however, the restrictions (8.208), (8.209) or (8.210) are not really crucial because the mesh coarseness is commonly not achieved and even if oscillations of this type are caused by too small time step sizes in a coarse mesh they are quickly damped out in progressing the time steps. Nevertheless, due to the higher accuracy the CM formulation (cf. Sect. 8.13.2) is generally the first choice in the present finite element analysis.

The stability criterion (8.198) represents a serious restriction for the explicit FE scheme ($\theta = 0$). A comprehensive stability analysis is given by Hindmarsh et al. [250]. The lumped explicit FE scheme becomes unstable, unless

$$Cr < \min(Pg, 1/Pg, 1) \tag{8.211}$$

or

$$\Delta t_n < \min\left(\frac{\acute{\mathcal{R}} \Delta x^2}{2D}, \frac{2\acute{\mathcal{R}} D}{q^2}, \frac{\acute{\mathcal{R}} \Delta x}{q}\right) \tag{8.212}$$

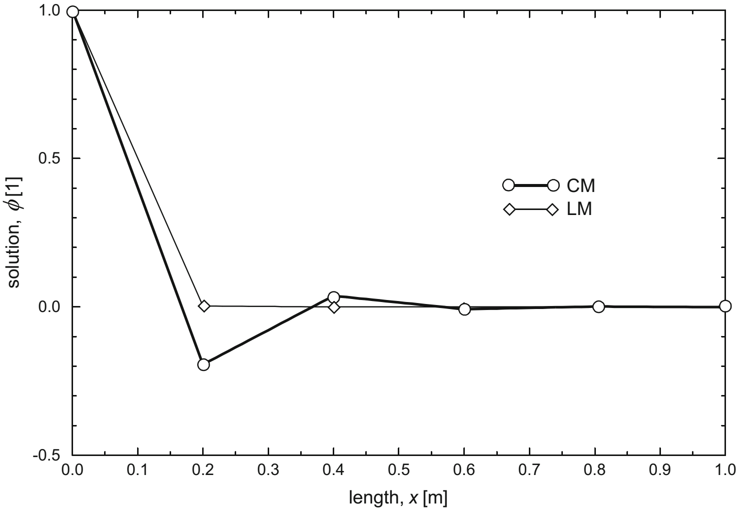


Fig. 8.22 Resulting distributions in a 1D domain of length $L = 1$ m discretized by five linear elements with $\Delta x = 0.2$ m, $D = 10^{-6}\,\mathrm{m^2\,s^{-1}}$, $q = 1\,\mathrm{m\,d^{-1}}$, $\acute{\mathcal{R}} = 1$ and $\vartheta = 3 \cdot 10^{-4}\,\mathrm{s^{-1}}$, satisfying the limit (8.209). Results are obtained for a full implicit scheme ($\theta = 1$) at the first time step of $\Delta t_n = 10^{-2}$ d for CM and LM formulations, where only CM implies bounded oscillations

The first 'diffusion limit' $\Delta t_n < \acute{\mathcal{R}} \Delta x^2/(2D)$ governs when $Pg < 1$, the second 'advection-diffusion limit' is restrictive when $Pg \geq 1$. The third restriction

$$Cr < 1 \tag{8.213}$$

is the *Courant-Friedrichs-Lewy (CFL) condition* [105], which is always a necessary condition for the stability of explicit schemes. If diffusion dominates the diffusion limit $\Delta t_n < \Delta x^2/(2D^*)$ possesses a terrible restriction for any explicit method. It means in practical terms: Assume L is the characteristic length of the computational domain, then the simulation time required for the full transient is $t_{\mathrm{end}} \simeq L^2/D^*$. For a typical (thermal) diffusivity D^* of $10^{-6}\,\mathrm{m^2\,s^{-1}}$, a length of $L = 10$ m and the smallest element length of $\Delta x = L/1{,}000 = 10^{-2}$ m, the diffusion limit (8.212) requires $\Delta t_n < 50$ s. Since $t_{\mathrm{end}} = 10^8$ s, about $2 \cdot 10^6$ time steps are required to perform the complete simulation with an explicit FE scheme. If we halve Δx the required times steps increase to $8 \cdot 10^6$ s. This shows the serious drawback of explicit schemes, in particular for diffusion problems, where a very large number of tiny time steps becomes necessary, albeit each time step is computationally cheap because no equation systems must be solved. In contrast, $\mathcal{A}$–stable implicit schemes having no stability limitations can solve a diffusion problem with acceptable accuracy in, let's say, less than 100 time steps, however, each time step is more expensive due to the solution of the equation system. Nevertheless, the implicit time stepping schemes have shown clearly superior to explicit methods, at least for diffusion-dominant problems, due to their clearly higher computational performance and robustness.

On the other hand, for dominant advection ($Pg \gg 1$), the CFL condition (8.213) becomes important for explicit schemes

$$\Delta t_n < \frac{\Delta x}{q^*} \tag{8.214}$$

which implies only a linear dependence on Δx. For example, choosing $L = 10\,\mathrm{m}$, $\Delta x = 10^{-2}\,\mathrm{m}$ and $q^* = 10^{-4}\,\mathrm{m\,s^{-1}}$, the time step limit (8.214) requires $\Delta t_n < 100\,\mathrm{s}$ and accordingly 10^3 time steps are needed to perform a full transient for the advection problem up to $t_{\mathrm{end}} \simeq L/q^* = 10^5\,\mathrm{s}$. It illustrates that the performance of explicit schemes considerably improves for advection-dominated (hyperbolic) problems (Sect. 8.3) and could be in fact more affordable compared to implicit techniques. But, taking into consideration more complex flow situations where locally (in space or time) the advection can be small compared to the diffusivity or even zero, the possible benefit of the computational performance of explicit methods can easily get lost again due to the strong limitation (8.212) in the time step control.

Finally, the advection-diffusion limit ($Cr < 1/Pg$ or $\Delta t_n < 2\acute{\mathcal{R}}D/q^2$) can be too restrictive for advection-dominated simulations via explicit methods [250]. However, it is common practice to incorporate the temporal truncation error and upwind stabilization techniques (see following Sect. 8.14), where the physical diffusion D^* is artificially increased by ${q^*}^2\Delta t_n/2$ and by $q^*\Delta x/2$, respectively [91,250]. Then, the explicit method becomes tractable for hyperbolic problems with the changed advection-diffusion limits according to [250]

$$Cr < \frac{\sqrt{1+4Pg^2}-1}{2Pg} \quad \text{and} \quad Cr < \frac{Pg}{1+Pg} \tag{8.215}$$

In a resumé, due to the desired generality and robustness of the finite element strategy we prefer usually implicit time stepping schemes for the present class of problems. Explicit techniques (such as FE and AB) only occur in the context of predictor-corrector time marching schemes, for which no time step restrictions exist because the corrector solutions are generally implicit in form of the $\mathcal{A}$−stable BE or TR methods.

8.14 Upwinding

8.14.1 *Pros and Cons of Upwind Methods*

In computing transport-flow processes the FEM must be applied in situations where the advection dominates over diffusion/dispersion. For the numerical solution stability and boundedness (definitions given in Sect. 1.2.2) should be guaranteed. Numerical solutions should lie within proper bounds. Physically, nonnegative quantities (e.g., density, mass concentration, absolute temperature) should always be positive. But, boundedness is difficult to guarantee under all circumstances. Unbounded solutions can occur on too coarse meshes in form of *wiggles*, i.e., oscillatory results generally occurring in a node-to-node manner which overshoot and undershoot the solution (Fig. 8.23). GFEM (and the equivalent central difference approximations) are prone to generate those spurious oscillations in space if the chosen mesh is inappropriate. In FDM it is popular to approximate the advective terms of the ADE by first-order onesided (flow direction-biased) differences, a process often referred to as *upwinding*. However, upwind methods precluding unwanted oscillations have disadvantages with regard to accuracy. It is to emphasize that stability does not imply accuracy – although it is true that instability implies inaccuracy. The resort to upwinding is usually a reduction of accuracy in favor of stability, where wiggles are artificially suppressed via damping mechanisms.

To treat advection-dominated transport problems by the FEM various upwind formulations have been developed in past. Pioneering work was given by Christie et al. [82], Heinrich et al. [236], Heinrich and Zienkiewicz [235] and Zienkiewicz et al. [595]. Asymmetric weighting functions were introduced such that the element upstream of a node is weighted more heavily than the element located downstream of a node equivalent to an upwind differencing. This type of upwind distortion of the weighting function represents a generalization of the standard (Galerkin-based) FEM and is called Petrov-Galerkin finite element method (PGFEM), cf. Table 8.1. Hughes [266] has shown that the upwind effect can also be achieved by asymmetry in the numerical quadrature rule for the advection terms. It was recognized that the PGFEM stabilization is equivalent to adding artificial (numerical) diffusion to the GFEM, termed as *balancing diffusion*, e.g., [307]. Unfortunately, many of the upwind methods reveal over-diffusive properties and there was a demand for alternative upwind techniques possessing reduced spurious numerical diffusion. While a scalar artificial diffusion often suffers from a considerable smearing effects [446, 541], the streamline-upwind (SU) method adds artificial diffusion only in the flow direction and not transversely [57]. The upwind finite element strategy have been further developed in a number of works, see e.g., [131, 149, 267–269, 272–276, 584, 585, 592]. The Petrov-Galerkin least square (PGLS) FEM [276, 385] appeared as a promising stabilization technique. This procedure results in an artificial diffusion concept of a built-in streamline-like upwinding similar to the SU method, however, leads to symmetric matrix systems. However, it has been found [274] that 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 [269]. The shock capturing (SC) method applied to finite elements has been developed by Johnson et al. [292] and Codina [90, 92, 93].

It becomes clear that upwinding is a compromise between the requirements of accuracy and stability. There is (also) ‘no free lunch’ in numerics: stability must be paid by a reduction of accuracy. The question arises how much reduction in accuracy is acceptable or to which level wiggling can be tolerated. The most important pros and cons of upwind methods can be summarized as follows:

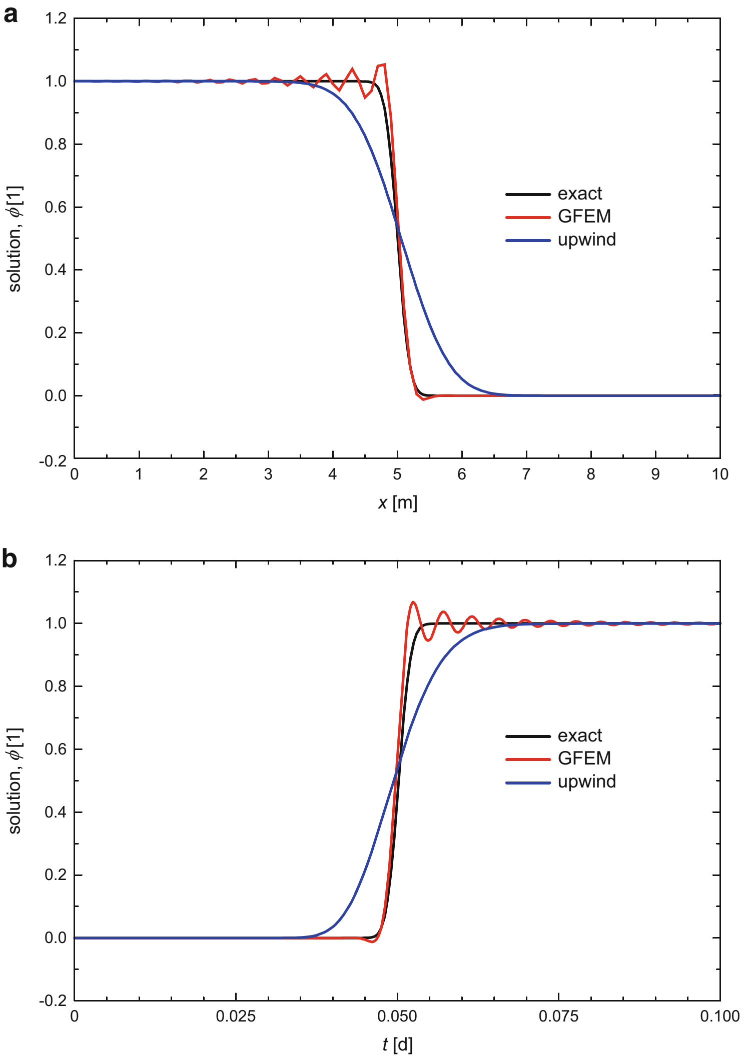


Fig. 8.23 (**a**) Profiles and (**b**) breakthrough curves for 1D advection-dominant transport in a uniform flow field obtained by GFEM and upwind method simulated with AB/TR predictor-corrector time stepping on a coarse mesh consisting of 100 linear elements with $\Delta x = 0.1\,\mathrm{m}$, $D^* = 2.5 \cdot 10^{-6}\,\mathrm{m^2\,s^{-1}}$ and $Pg = 23.15$ in comparison to the exact solution.[12] Oscillations are generated for GFEM, while smooth and overdiffusive solution results for upwinding, where physical diffusion D^* is artificially increased to $D^* + q^* \Delta x/2 = 6.04 \cdot 10^{-5}\,\mathrm{m^2\,s^{-1}}$, which is more than 24 times higher

[12] The analytical (exact) solution of a 1D ADE is [71], p. 388, [540], cf. also Sect. 12.5.1

Pros

- GFEM has serious deficiencies in solving problems with dominant advection, which are prone to generate spurious (nonphysical) node-to-node oscillations. Upwinding can *stabilizes* the solutions and is beneficial to obtain realistic (though not always accurate) solutions.
- Upwind methods allow *efficient* solutions without the ultimate need for fine (sometimes extremely dense) meshes.
- Upwinding makes difficult problems *computable* under given computational constraints. Extremely fine meshes and expensive computations could be caused for tough physical situations (e.g., shock-like front displacements of mass or energy, very thin boundary layers, high density contrasts in a large-scale problem) if upwind methods would not be admitted.
- There are certain situations where any wiggles in the solution become absolutely devastating and would totally preclude the possibility of obtaining a solution, e.g., strong advection in multispecies mass transport processes with nonlinear chemical reaction.

Cons

- 'Don't suppress the wiggles – they're telling you something!' as stated in the famous paper by Gresho and Sani [208] who oppose, in principle, any artificial damping measures by upwinding: Wiggles are usually a signal that the spatial (and temporal) discretization is poor and some mesh refinements (at least locally) are required to obtain a physically adequate solution.
- A positive aspect of wiggles is that in signaling improper discretization they present *self-diagnosis property*. A method with such a self-diagnostic property is often superior to schemes which give smooth, and totally wiggle-free, but inaccurate and possibly overdamped solutions for any discretization.
- Upwinding is a method of damping and smoothing. *It solves the problem by changing the physics of the problem.* Robustness is obtained at the expense of accuracy. Diffusion is artificially increased in dependence on the chosen mesh, i.e., the solution becomes mesh-dependent. With other words: For a coarse mesh the solution is independent of the physical diffusion and can be depart from the physics of the original problem. Upwinding could be only acceptable

$$\phi(x,t) = \phi_0 + \tfrac{1}{2}(\phi_D - \phi_0)\left[\operatorname{erfc}\left(\frac{x - q^*t}{2\sqrt{D^*t}}\right) + \exp\left(\frac{xq^*}{D^*}\right)\operatorname{erfc}\left(\frac{x + q^*t}{2\sqrt{D^*t}}\right)\right]$$

valid for the IC: $\phi(x,0) = \phi_0$, and BC's: $\phi(0,t) = \phi_D$ and $\frac{\partial\phi}{\partial x}(\infty,t) = 0$, where

$$\operatorname{erfc}(a) = \tfrac{2}{\sqrt{\pi}}\int_a^\infty \exp(-\xi^2)d\xi$$

is the complementary error function [71], $\phi_0 = 0$ is the used initial value and $\phi_D = 1$ is the used Dirichlet-type BC at $x = 0$. Note that for evaluating the analytical exp(.)erfc(.) expression the more suitable exf(., .) function is applied which will be further discussed in Sect. 12.5.1.

and reasonably accurate if the numerical diffusion is significantly less than the physical diffusion.

- Upwinding is potentially dangerous because it often leads to a false sense of security: 'Any mesh works for any advection-diffusion relation'. Upwind schemes can damp more than just wiggles; this is particularly true for more complex nonlinear problems. J. Ferziger (noted in [209]) stated: *The greatest disaster one can encounter in computation is not instability or lack of convergence, but results that are good enough to be believable but bad enough to cause trouble*.

The decision between the pros and cons is often not easy. There is, unfortunately, no panacea, but the practitioner should be aware of the necessary compromises involved, and use a given method with due caution and 'healthy skepticism'. Finally, our recommended strategy is to solve a problem without upwinding whenever possible, and to resort to an upwind method only if necessary and unavoidable. In the following, appropriate upwind methods available in FEFLOW will be described.

8.14.2 Petrov-Galerkin Finite Element Method (PGFEM)

The most common technique for introducing the upwind concept into the FEM is the Petrov-Galerkin finite element method (PGFEM), where the element weighting functions differ from the element basis functions $w_I^e \neq N_I^e$ (cf. Table 8.1) and are appropriately designed to incorporate asymmetry with respect to the flow field. The weighting functions of an element e are constructed in general as

$$w_I^e(\boldsymbol{\eta}) = N_I^e(\boldsymbol{\eta}) + \alpha F_I^e(\boldsymbol{\eta}) \tag{8.216}$$

where F_I^e are modifying functions with the sign depending on the sign of the advective flux $\boldsymbol{q}$ and α is a free, so-called *upwind parameter* ($0 \leq \alpha \leq 1$), which has to be determined. We note if $\alpha = 0$, (8.216) corresponds to the standard GFEM. The modifying functions F_I^e can be appropriately chosen either as continuous and discontinuous relations. Let us consider for convenience firstly the 1D case: In the continuous definition F_I^e are chosen as a polynomial one degree higher than N_I^e, e.g., [236]. For a linear 1D element we introduce $F_I^e(\xi) = \mp f^e(\xi)$, where $f^e(\xi) = a\xi^2 + b\xi + c$, written in the local coordinate $(-1 \leq \xi \leq +1)$, and determine its polynomial coefficients a, b and c such that $f^e(-1) = f^e(1) = 0$ and $\int_{-1}^{+1} f^e(\xi)d\xi = 1$. It leads to

$$f^e(\xi) = \tfrac{3}{4}(1-\xi)(1+\xi) \tag{8.217}$$

Then, the following continuous weighting functions result as shown in Fig. 8.24 for a linear 1D element at the local nodes 1 and 2:

$$\begin{aligned} w_1^e(\xi) &= N_1^e(\xi) - \alpha f^e(\xi) \\ w_2^e(\xi) &= N_2^e(\xi) + \alpha f^e(\xi) \end{aligned} \tag{8.218}$$

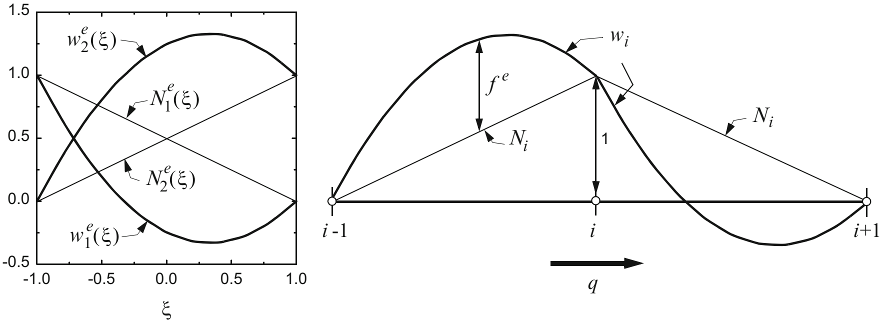


Fig. 8.24 Continuous Petrov-Galerkin weighting functions ($\alpha = 1$) for the linear 1D element

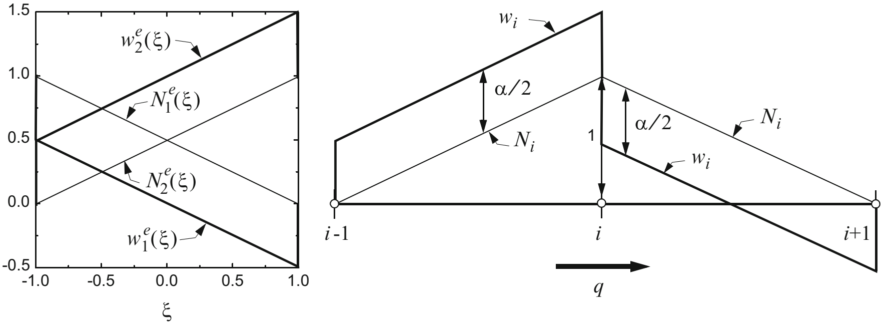


Fig. 8.25 Discontinuous Petrov-Galerkin weighting functions ($\alpha = 1$) for the linear 1D element

where the basis functions are $N_1^e(\xi) = \frac{1}{2}(1 - \xi)$, $N_2^e(\xi) = \frac{1}{2}(1 + \xi)$, cf. (H.1) of Appendix H.

Much more convenient are discontinuous weighting functions, e.g., [584]. In 1D one simply chooses:

$$\alpha F_I^e(\xi) = \alpha \frac{\Delta x^e}{2} \frac{q^e}{|q^e|} \frac{dN_I^e}{dx} \tag{8.219}$$

where Δx^e is the length of the finite element e. By using the derivations (H.6) of Appendix H we obtain the following asymmetric discontinuous weighting functions for a linear 1D element at the local nodes 1 and 2 with a positive advective flux $q^e > 0$

$$\begin{aligned} w_1^e(\xi) &= N_1^e(\xi) - \tfrac{\alpha}{2} \\ w_2^e(\xi) &= N_2^e(\xi) + \tfrac{\alpha}{2} \end{aligned} \tag{8.220}$$

which are displayed in Fig. 8.25.

It is important to note that the discontinuous weighting functions w_I^e must result finite contributions for first derivatives in the integrand of the approximate weak statement in order satisfy the requirement on continuity as stated in Sect. 8.7. The discontinuity of the formulations (8.219) or (8.220) is considered within the element such that any first-order derivative of w_I^e is finite and valuable due to the C_0-continuity in N_I^e.

Usually, the asymmetric weighting functions w_I^e are only applied to the terms of the homogeneous solution of the governing PDE, i.e., in particular the advection and diffusion/dispersion terms. In doing so, for example for the 1D discrete finite element equations of Sect. H.1 of Appendix H, we find for (H.10) a modified formulation of the semidiscrete ADE convective form[13] (for sake of simplicity we drop BC and SPC terms):

$$\begin{aligned}\sum_e \Bigg\{ & \frac{\acute{\mathcal{R}}^e \Delta x^e}{6}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\cdot\begin{pmatrix}\frac{d\phi_1^e}{dt}\\ \frac{d\phi_2^e}{dt}\end{pmatrix} + \Bigg[\underbrace{\frac{q^e}{2}\begin{pmatrix}-1+\alpha & 1-\alpha\\ -1-\alpha & 1+\alpha\end{pmatrix}}_{\boldsymbol{A}^e} + \underbrace{\frac{D^e}{\Delta x^e}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}}_{\boldsymbol{C}^e} \\ & + \frac{(\vartheta^e + Q^e)\Delta x^e}{6}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\Bigg]\cdot\begin{pmatrix}\phi_1^e\\ \phi_2^e\end{pmatrix} - \frac{H^e \Delta x^e}{2}\begin{pmatrix}1\\ 1\end{pmatrix}\Bigg\} = \mathbf{0}\end{aligned} \tag{8.221}$$

We recognize from (8.221) that the upwind parameter α is indeed only effective in the advection matrix $\boldsymbol{A}^e$, while it is canceled out in the diffusion matrix $\boldsymbol{C}^e$. Furthermore, it is easy to see that the sum of $\boldsymbol{A}^e + \boldsymbol{C}^e$ can be alternatively written as[14]

$$\boldsymbol{A}^e + \boldsymbol{C}^e = \frac{q^e}{2}\begin{pmatrix}-1 & 1\\ -1 & 1\end{pmatrix} + (D^e + \alpha\frac{q^e \Delta x^e}{2})\frac{1}{\Delta x^e}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix} \tag{8.222}$$

where the physical diffusion D^e is increased by $\alpha\frac{q^e \Delta x^e}{2}$, which represents the artificial diffusion introduced by the PGFEM upwind method. Similar to (8.203) the assembly of the linear 1D elements (8.221) and applying the temporal $\theta-$integration scheme of (8.154) the following discrete equations of the 3-node $(i+1, i, i-1)-$stencil can be written assuming constant parameter properties (and for convenience also dropping source/sink terms):

[13] For the ADE convective form the continuous weighting functions (8.218) and the discontinuous weighting functions (8.220) lead to the same result. However, for the ADE divergence form only the continuous weighting functions (8.218) are applicable, where the element advection matrix $\boldsymbol{A}^e$ (8.104) becomes

$$\boldsymbol{A}^e = \frac{q^e}{2}\begin{pmatrix}1+\alpha & 1-\alpha\\ -1-\alpha & -1+\alpha\end{pmatrix}$$

[14] A similar expression can be obtained for the ADE divergence form.

$$\begin{aligned}&\Big[\frac{c}{6}-\Delta t_n\theta\Big(1+(\alpha-1)Pg\Big)\Big]\phi_{i+1,n+1}+\Big[\frac{2c}{3}+\Delta t_n\theta\Big(2+2\alpha Pg\Big)\Big]\phi_{i,n+1}+\\ &\Big[\frac{c}{6}-\Delta t_n\theta\Big(1+(\alpha+1)Pg\Big)\Big]\phi_{i-1,n+1}=\Big[\frac{c}{6}+\Delta t_n(1-\theta)\Big(1+(\alpha-1)Pg\Big)\Big]\phi_{i+1,n}+\\ &\Big[\frac{2c}{3}-\Delta t_n(1-\theta)\Big(2+2\alpha Pg\Big)\Big]\phi_{i,n}+\Big[\frac{c}{6}+\Delta t_n(1-\theta)\Big(1+(\alpha+1)Pg\Big)\Big]\phi_{i-1,n}\end{aligned} \tag{8.223}$$

where $c = \Delta x^2/D^*$, $D^* = D/\acute{\mathcal{R}}$ and Pg is the grid Réclet number as defined in (8.207).

To analyze the upwind parameter α let us at first turn to the steady-state ADE formulation. Then, (8.223) reduces to

$$[1+(\alpha-1)Pg]\phi_{i+1}-2(1+\alpha Pg)\phi_i+[1+(\alpha+1)Pg]\phi_{i-1}=0 \tag{8.224}$$

which represents a PGFEM formulation of the simplified 1D ADE:

$$q\nabla\phi - D\nabla^2\phi = 0 \quad \text{where in 1D} \quad \nabla = \partial/\partial x \tag{8.225}$$

We can solve (8.225) within the interval $x_{i-1} \leq x \leq x_{i+1}$ for the local boundary value problem: $\phi(x_{i-1}) = \phi_{i-1}$ and $\phi(x_{i+1}) = \phi_{i+1}$. The exact solution of this local 1D problem is

$$\phi(x)=\phi_{i-1}+\big(\phi_{i+1}-\phi_{i-1}\big)\frac{\exp[\frac{2Pg}{\Delta x}(x-x_{i-1})]-1}{\exp(4Pg)-1} \tag{8.226}$$

which can be taken to express the solution for $\phi(x_i) = \phi_i$ leading to the 'locally-exact' formula of the 3-node $(i+1, i, i-1)$−stencil

$$\phi_{i+1}-(1+a)\phi_i+a\phi_{i-1}=0 \tag{8.227}$$

where

$$a = \exp(2Pg) \tag{8.228}$$

with

$$\begin{array}{lll} a > 0 & \text{for} & q > 0 \\ \frac{1}{a} > 0 & \text{for} & q < 0 \end{array} \tag{8.229}$$

In comparison of the scheme (8.224) with the exact formula (8.227) it must be required due to (8.229) for $q > 0$

$$a = \frac{1 + Pg(\alpha + 1)}{1 + Pg(\alpha - 1)} > 0 \tag{8.230}$$

which yields

$$\begin{array}{lll} \alpha \geq \alpha^{\text{crit}} = 1 - \frac{1}{Pg} & \text{for} & Pg \geq 1 \\ \alpha = 0 & \text{for} & Pg < 1 \end{array} \tag{8.231}$$

Apparently, for the standard GFEM with $\alpha = 0$ once the grid Péclet number $Pg > 1$ node-to-node oscillations will occur since the denominator in (8.230) becomes negative. To avoid oscillations the upwind parameter α must be greater than the critical value α^{crit} defined in (8.231). It shows that for $\alpha = 1$ the scheme is unconditionally stable and corresponds to a *full upwind* scheme. Furthermore, a complete accuracy is obtained for a given Péclet number Pg if the parameter a of the exact solution (8.228) is equated to a of the approximate solution (8.230). It gives the so-called *optimal upwind parameter*

$$\alpha^{\text{opt}} = \coth(Pg) - \frac{1}{Pg} \tag{8.232}$$

It is obvious that the α^{opt} satisfies the stability criterion (8.231) with

$$\alpha^{\text{opt}} \geq \alpha^{\text{crit}} \tag{8.233}$$

and so, indeed, it is optimal for this class of problems (Fig. 8.26). Upwind parameter relations for higher-order finite elements have been derived in [81, 131, 235].

The extension of the PGFEM to multidimensional and transient ADE problems can be done straightforward, e.g., [278, 279, 592]. However, in 2D and particularly in 3D the use of continuous weighting functions in form of (8.216) is cumbersome and ineffective, so that discontinuous weighting functions are often preferred. An appropriate discontinuous weighting function is [585]

$$w_I^e(\boldsymbol{\eta}) = N_I^e(\boldsymbol{\eta}) + \frac{h^e}{2\|\boldsymbol{q}^e\|}\left(\alpha + \frac{\beta \Delta t_n}{2}\frac{\partial}{\partial t}\right)\boldsymbol{q}^e \cdot \nabla N_I^e(\boldsymbol{\eta}) \tag{8.234}$$

where h^e is a characteristic element length which is defined further below, α is a first upwind parameter as already introduced above for steady-state problems and β is a second upwind parameter related to the transient terms of the ADE. The intent and result of (8.234) is to add artificial diffusion into the discrete finite element equations. With the upwind parameter α the diffusion is increased by $\alpha\|\boldsymbol{q}^e\|h^e/2$ and with the upwind parameter β an added diffusion term is in the order of $\beta\|\boldsymbol{q}^e\|h^e\Delta t_n/4$. The upwind parameters can be determined by Yu and Heinrich [584]

$$\begin{array}{l} \alpha = \coth(Pg) - \frac{1}{Pg} \\ \beta = \frac{Cr}{3} - \frac{\alpha}{PgCr} \end{array} \tag{8.235}$$

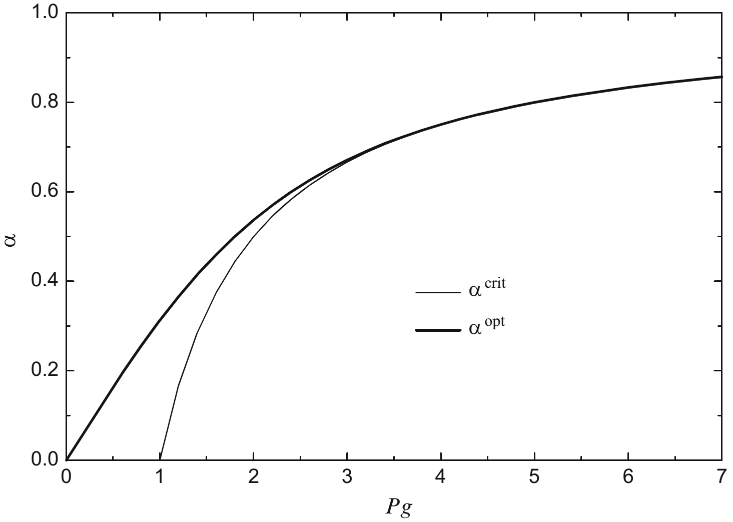


Fig. 8.26 Critical upwind parameter α^{crit} and optimal upwind parameter α^{opt} in dependence on the grid Péclet number Pg based on linear finite elements

where the mesh Péclet number Pg (8.207) and the Courant number Cr (8.206) are defined in multidimensions

$$Pg = \frac{\|\boldsymbol{q}^{*e}\| h^e}{2D^{*e}}, \quad Cr = \frac{\|\boldsymbol{q}^{*e}\| \Delta t_n}{h^e} \quad \text{with} \quad \|\boldsymbol{q}^{*e}\| = \|\boldsymbol{q}^e\|/\acute{\mathcal{R}}, \quad D^{*e} = D^e/\acute{\mathcal{R}} \tag{8.236}$$

The resulting PGFEM upwind scheme [584, 585] shows the best accuracy with $\alpha \neq 0$ and $\beta \neq 0$ according to (8.235). However, unconditionally stable algorithms also result for $\alpha \neq 0$ and $\beta = 0$, albeit more artificial diffusion is produced.

The PGFEM upwind scheme in multidimensions requires the determination of the characteristic element length h^e. Figure 8.27 shows typical isoparametric finite elements in 2D and 3D over which the parametric vectors $\boldsymbol{h}_\xi$, $\boldsymbol{h}_\eta$ and $\boldsymbol{h}_\zeta$ are defined and computed in 2D as

$$\left.\begin{aligned} \boldsymbol{h}_\xi = h_{\xi i} &= \tfrac{1}{2}[(x_{i2} + x_{i3}) - (x_{i1} + x_{i4})] \\ \boldsymbol{h}_\eta = h_{\eta i} &= \tfrac{1}{2}[(x_{i3} + x_{i4}) - (x_{i1} + x_{i2})] \end{aligned}\right\} \quad i = 1,2 \tag{8.237}$$

and in 3D as

$$\left.\begin{aligned} \boldsymbol{h}_\xi = h_{\xi i} &= \tfrac{1}{4}[(x_{i2} + x_{i3} + x_{i6} + x_{i7}) - (x_{i1} + x_{i4} + x_{i5} + x_{i8})] \\ \boldsymbol{h}_\eta = h_{\eta i} &= \tfrac{1}{4}[(x_{i3} + x_{i4} + x_{i7} + x_{i8}) - (x_{i1} + x_{i2} + x_{i5} + x_{i6})] \\ \boldsymbol{h}_\zeta = h_{\zeta i} &= \tfrac{1}{4}[(x_{i1} + x_{i2} + x_{i3} + x_{i4}) - (x_{i5} + x_{i6} + x_{i7} + x_{i8})] \end{aligned}\right\} \quad i = 1,2,3 \tag{8.238}$$

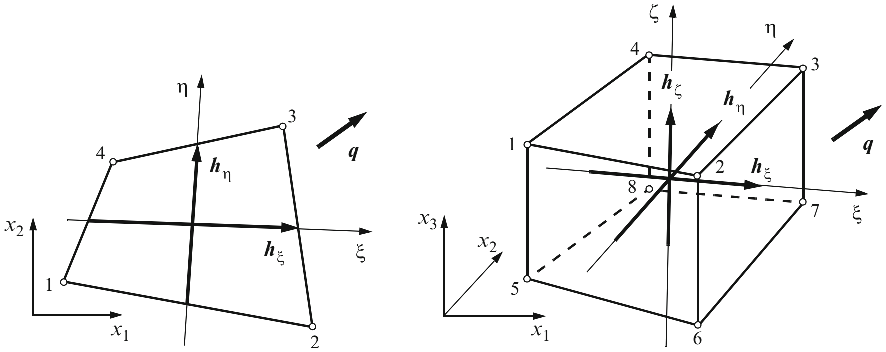


Fig. 8.27 2D quadrilateral and 3D hexahedral element used in definition of element length h^e

Similar relations can be obtained for the other finite elements listed in Tables G.2–G.4 of Appendix G. The characteristic element length h^e then results

$$h^e = \begin{cases} |h_1| + |h_2| & \text{for 2D} \\ |h_1| + |h_2| + |h_3| & \text{for 3D} \end{cases} \tag{8.239}$$

where in 2D

$$\begin{aligned} h_1 &= \frac{1}{\|\boldsymbol{q}^e\|}(\boldsymbol{q}^e \cdot \boldsymbol{h}_\xi) = \frac{1}{\|\boldsymbol{q}^e\|}(q_1^e h_{\xi 1} + q_2^e h_{\xi 2}) \\ h_2 &= \frac{1}{\|\boldsymbol{q}^e\|}(\boldsymbol{q}^e \cdot \boldsymbol{h}_\eta) = \frac{1}{\|\boldsymbol{q}^e\|}(q_1^e h_{\eta 1} + q_2^e h_{\eta 2}) \end{aligned} \tag{8.240}$$

and in 3D

$$\begin{aligned} h_1 &= \frac{1}{\|\boldsymbol{q}^e\|}(\boldsymbol{q}^e \cdot \boldsymbol{h}_\xi) = \frac{1}{\|\boldsymbol{q}^e\|}(q_1^e h_{\xi 1} + q_2^e h_{\xi 2} + q_3^e h_{\xi 3}) \\ h_2 &= \frac{1}{\|\boldsymbol{q}^e\|}(\boldsymbol{q}^e \cdot \boldsymbol{h}_\eta) = \frac{1}{\|\boldsymbol{q}^e\|}(q_1^e h_{\eta 1} + q_2^e h_{\eta 2} + q_3^e h_{\eta 3}) \\ h_3 &= \frac{1}{\|\boldsymbol{q}^e\|}(\boldsymbol{q}^e \cdot \boldsymbol{h}_\zeta) = \frac{1}{\|\boldsymbol{q}^e\|}(q_1^e h_{\zeta 1} + q_2^e h_{\zeta 2} + q_3^e h_{\zeta 3}) \end{aligned} \tag{8.241}$$

are the projections of $\boldsymbol{h}_\xi$, $\boldsymbol{h}_\eta$ and $\boldsymbol{h}_\zeta$ in the direction of the local flow vector $\boldsymbol{q}^e$. We note that for rectangular geometries the expression (8.239) reduces to $h^e = (|q_1^e|\Delta x_1^e + |q_2^e|\Delta x_2^e + |q_3^e|\Delta x_3^e)/\|\boldsymbol{q}^e\|$ in 3D, where Δx_i^e, $(i = 1, 2, 3)$ are the lengths of element edges in the coordinate directions. In 1D geometries it is simply $h^e = \Delta x^e$.

8.14.3 Streamline Upwind (SU) and Full Upwind (FU) Method

We have seen in the preceding Sect. 8.14.2 that PGFEM is designed to add an appropriate amount of artificial diffusion for stabilization purposes. The use of the asymmetric weighting function takes effect only on the advective term and ends up with a diffusion increased for instance by $\alpha\|\boldsymbol{q}^e\|h^e/2$ for a linear finite element. It is obvious that such a type of stabilization should be correlated with the flow direction only and should not be effective in the transverse direction of advection to avoid an overly diffusion due to an excess of so-called *crosswind diffusion*.

The avoidance of crosswind diffusion leads to the concept of the streamline upwind (SU) method. The basic ideas were given by Kelly et al. [307] and Brooks and Hughes [57] who constructed the artificial diffusion operator in tensorial form acting only in the flow direction and not transversely, termed as *anisotropic balancing dissipation*. The idea of SU is to extend the tensor of physical dispersion/diffusion $\boldsymbol{D}$ defined for example in a porous medium of a single-species solute transport as (cf. (3.180), Tables 3.7 and 3.9)

$$\begin{aligned} \boldsymbol{D} &= \varepsilon s D\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}} \\ \boldsymbol{D}_{\text{mech}} &= \beta_T\|\boldsymbol{q}\|\boldsymbol{\delta} + (\beta_L - \beta_T)\frac{\boldsymbol{q}\otimes\boldsymbol{q}}{\|\boldsymbol{q}\|} \end{aligned} \tag{8.242}$$

where $\boldsymbol{D}_{\text{mech}}$ is the (physical) tensor of mechanical dispersion and D is the molecular diffusion, by the tensor of numerical dispersion $\boldsymbol{D}_{\text{num}}$ in the form

$$\boldsymbol{D}_{\text{num}} = \beta_{\text{num}}\frac{\boldsymbol{q}\otimes\boldsymbol{q}}{\|\boldsymbol{q}\|} \tag{8.243}$$

so that

$$\boldsymbol{D} = \varepsilon s D\boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}} + \boldsymbol{D}_{\text{num}} = \varepsilon s D\boldsymbol{\delta} + \beta_T\|\boldsymbol{q}\|\boldsymbol{\delta} + (\beta_L + \beta_{\text{num}} - \beta_T)\frac{\boldsymbol{q}\otimes\boldsymbol{q}}{\|\boldsymbol{q}\|} \tag{8.244}$$

where β_{num} represents the parameter of *numerical longitudinal dispersivity* which must be specified for each element. For example, in case of linear elements one takes for the element e

$$\beta_{\text{num}}^e = \alpha\frac{h^e}{2} \tag{8.245}$$

where $0 \leq \alpha \leq 1$ is the upwind parameter introduced above ($\alpha = 0$ is the standard GFEM, $\alpha = 1$ is the full upwind, $\alpha = \alpha^{\text{opt}}$ is the optimal parameter defined in (8.232)) and h^e is the characteristic element length defined in (8.239). Note that for quadratic elements $\beta_{\text{num}}^e = \alpha h^e/4$ as derived in [131].

Now, if looking to the resulting weak statement we have to modify for the advective and dispersive terms of the governing ADE written in its convective form according to (8.55)

$$\text{WS} = \int_\Omega w\boldsymbol{q}\cdot\nabla\phi\, d\Omega + \int_\Omega \nabla w\cdot[(\boldsymbol{D} + \boldsymbol{D}_{\text{num}})\cdot\nabla\phi]d\Omega \tag{8.246}$$

Since

$$\int_{\Omega} \nabla w \cdot (\boldsymbol{D}_{\text{num}} \cdot \nabla \phi) d\Omega = \int_{\Omega} \tfrac{\beta_{\text{num}}}{\|\boldsymbol{q}\|} (\boldsymbol{q} \cdot \nabla w)(\boldsymbol{q} \cdot \nabla \phi) d\Omega \tag{8.247}$$

Eq. (8.246) can be rewritten as

$$\text{WS} = \int_{\Omega} [w + \tfrac{\beta_{\text{num}}}{\|\boldsymbol{q}\|} (\boldsymbol{q} \cdot \nabla w)](\boldsymbol{q} \cdot \nabla \phi) d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla \phi) d\Omega \tag{8.248}$$

As a result, a modified SU weighting function can be found in the form

$$\tilde{w} = w + \tfrac{\beta_{\text{num}}}{\|\boldsymbol{q}\|} (\boldsymbol{q} \cdot \nabla w) \tag{8.249}$$

which only affects the advective term and is similar to the discontinuous weighting function (8.234) (for $\beta = 0$) used by the PGFEM. Finally, the SU method is recognized as the standard GFEM plus an extra term introducing the SU added numerical dispersion term:

$$\text{GWS} = \underbrace{\sum_e \int_{\Omega^e} [N_i \boldsymbol{q} \cdot \nabla \phi + \nabla N_i \cdot (\boldsymbol{D} \cdot \nabla \phi)] d\Omega^e}_{\text{standard GFEM}} + \underbrace{\sum_e \int_{\Omega^e} \tfrac{\beta^e_{\text{num}}}{\|\boldsymbol{q}\|} (\boldsymbol{q} \cdot \nabla N_i)(\boldsymbol{q} \cdot \nabla \phi) d\Omega^e}_{\text{added SU stabilization term}} \tag{8.250}$$

where $\boldsymbol{D}$ represents the physical dispersion tensor (8.242), $\beta^e_{\text{num}} = \alpha h^e/2$ for linear elements and $\beta^e_{\text{num}} = \alpha h^e/4$ for quadratic elements. In practice, however, the second SU stabilization term in (8.250) in not directly executed. Instead, the modified dispersion tensor (8.244) is employed in the standard GFEM term, which is equivalent to (8.250).

Commonly, the SU method is used with $\alpha = 1$. In case of need the SU stabilization can be turned back to a *full upwinding* (FU), where the stabilization is performed in all coordinate directions, i.e., independent of the flow field. In the full upwind case the dispersion tensor (8.244) is then used in the form:

$$\boldsymbol{D} = \varepsilon s D \boldsymbol{\delta} + (\beta_T + \beta_{\text{num}}) \|\boldsymbol{q}\| \boldsymbol{\delta} + (\beta_L - \beta_T) \tfrac{\boldsymbol{q} \otimes \boldsymbol{q}}{\|\boldsymbol{q}\|} \tag{8.251}$$

However, it should be aware that a full upwind scheme usually produces a large amount of crosswind diffusion.

8.14.4 Shock Capturing (SC) Method

SU stabilization is only effective in longitudinal direction of the advective flow and avoids any crosswind damping. This is motivated by the fact that often the gradient of a transported quantity ϕ establishes in the direction of flow. However, under

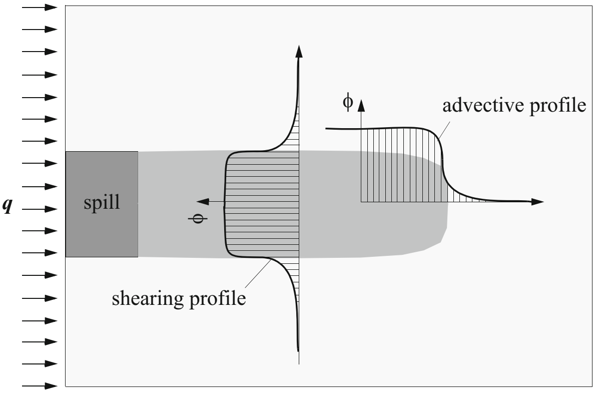


Fig. 8.28 Solution profiles $\phi(\boldsymbol{x}, t)$ in longitudinal and transverse direction to a flow field $\boldsymbol{q}$ forming an advection-displaced front and shearing transition layers, respectively

more complex flow conditions steep gradients can also occur in directions normal or skewed to the advective flow forming shearing (or flushing) transition layers such as illustrated in Fig. 8.28. Then, oscillations cannot be stabilized via the SU method. Indeed, it has been shown by Hughes et al. [274] that the streamline is not always the appropriate upwind direction. To increase the robustness of upwind methods it is obvious that the control of gradients must be required, i.e., the upwind direction should be aligned to the direction of gradients $\nabla\phi$ of the transported quantity ϕ rather than exclusively oriented to the trajectory of flow. This was basically proposed by Hughes and Mallet [269] who generalized the SU concept by adding an additional diffusion in the gradient direction which is called *discontinuity capturing* or *shock capturing* (SC).

The SC method has been further developed by Johnson et al. [292] and Codina [90]. The SC technique appears as a nonlinear method because the gradient $\nabla\phi$ is part of the numerical solution. The main idea behind SC is to increase the amount of damping in the neighborhood of gradients. Then, the damping to be added must be proportional to the discrete residual of the governing ADE within each element and must be vanish in regions where the solution is smooth and also where the advective term of the residual is small. Hence, SC stabilizes in dependence on the solution gradient and is accordingly operational both in longitudinal and transverse direction. It admits an optimal amount of crosswind damping necessary to stabilize also the shearing profiles (Fig. 8.28).

We have shown in the preceding Sect. 8.14.3 that the SU method is characterized by introducing an additional term $\frac{\beta_{\text{num}}}{\|\boldsymbol{q}\|}(\boldsymbol{q} \cdot \nabla w)$ to the weighting function $\tilde{w}$ in form of (8.249). Now, the basic idea of SC is to use $\tilde{w}$ with a further additional term, the SC term, such that

$$\tilde{w} = \underbrace{w}_{\text{GFEM}} + \underbrace{\tau_1(\boldsymbol{q} \cdot \nabla w)}_{\text{SU}} + \underbrace{\tau_2(\boldsymbol{q}_{\|} \cdot \nabla w)}_{\text{SC}} \tag{8.252}$$

where the first term is the standard Galerkin weighting function, the second term is the linear SU modification and the third term is the new nonlinear SC extension. The vector $\boldsymbol{q}_{\parallel}$ is the projection of the flux vector $\boldsymbol{q}$ onto the direction of the local gradient $\nabla\phi$ of the solution ϕ, viz.,

$$\boldsymbol{q}_{\parallel} = \frac{\boldsymbol{q} \cdot \nabla\phi}{\|\nabla\phi\|^2} \nabla\phi \tag{8.253}$$

provided that $\|\nabla\phi\| \neq 0$. It is easy to see that $\boldsymbol{q}_{\parallel} \cdot \nabla\phi = \boldsymbol{q} \cdot \nabla\phi$. The upwind parameters τ_1 and τ_2 are defined on element level as

$$\tau_1 = \frac{\alpha h^e}{2\|\boldsymbol{q}\|}, \quad \tau_2 = \max\left(0, \frac{\alpha h^e}{2\|\boldsymbol{q}_{\parallel}\|} - \tau_1\right) \tag{8.254}$$

written for linear elements, where $0 \leq \alpha \leq 1$ is the known upwind parameter defined above, (8.231) or (8.232), and h^e is the characteristic element length according to (8.239). In using (8.252), the SC method is recognized as the standard GFEM plus two extra terms introducing the SU added numerical dispersion term and the SC added numerical dispersion term applied to the Galerkin weak statement of the advective and dispersive terms of the governing ADE:

$$\begin{aligned} \text{GWS} = & \underbrace{\sum_e \int_{\Omega^e} [N_i \boldsymbol{q} \cdot \nabla\phi + \nabla N_i \cdot (\boldsymbol{D} \cdot \nabla\phi)] d\Omega^e}_{\text{standard GFEM}} + \underbrace{\sum_e \int_{\Omega^e} \tau_1 (\boldsymbol{q} \cdot \nabla N_i)(\boldsymbol{q} \cdot \nabla\phi) d\Omega^e}_{\text{added SU stabilization term}} \\ & + \underbrace{\sum_e \int_{\Omega^e} \tau_2 (\boldsymbol{q}_{\parallel} \cdot \nabla N_i)(\boldsymbol{q}_{\parallel} \cdot \nabla\phi) d\Omega^e}_{\text{added SC stabilization operator}} \end{aligned} \tag{8.255}$$

where the SC method is constructed to keep unaltered the added numerical dispersion in the streamline direction and to modify only the crosswind (transverse) dispersion. This crosswind dispersion must satisfy two conditions [90]. First, to avoid overdamped crosswind effects, it must be small in regions where the advective transport is not important, that is where $\boldsymbol{q} \cdot \nabla\phi$ is small. Second, the measure of crosswind damping should be proportional to the element residual, e.g., for the ADE convective form

$$R(\phi) = \boldsymbol{q} \cdot \nabla\phi - \nabla \cdot (\boldsymbol{D} \cdot \nabla\phi) + (\vartheta + Q)\phi - H - Q_{\phi w} \tag{8.256}$$

to be evaluated on element basis. Using $R(\phi)$ we can determine an isotropic SC dispersion coefficient as [90, 292]

$$D_{\mathrm{sc}} = \tfrac{1}{2}\alpha_c h^e \frac{|R(\phi)|}{\|\nabla\phi\|} \tag{8.257}$$

if $\|\nabla\phi\| \neq 0$ and zero otherwise. If we use the element residual only for the advective term of the ADE convective form by $R(\phi) \approx \boldsymbol{q}\cdot\nabla\phi$, a useful and simplified estimate of the isotropic SC dispersion coefficient results

$$D_{\mathrm{sc}} \approx \tfrac{1}{2}\alpha_c h^e \|\boldsymbol{q}_{\shortparallel}\| \tag{8.258}$$

where with (8.253) it is $\|\boldsymbol{q}_{\shortparallel}\| = |\boldsymbol{q}\cdot\nabla\phi|/\|\nabla\phi\|$. The upwind parameter α_c is given by

$$\alpha_c = \max\Big(0, a - \frac{1}{Pg_{\shortparallel}}\Big) \quad \text{with} \quad Pg_{\shortparallel} = \frac{\|\boldsymbol{q}_{\shortparallel}\| h^e}{2D^e} \tag{8.259}$$

where it is proposed, e.g., [90]

$$a = \begin{cases} \left.\begin{array}{ll} 0.7 & \text{for linear element} \\ 0.35 & \text{for quadratic element} \end{array}\right\} & \text{in 2D} \\ \left.\begin{array}{ll} 1.0 & \text{for linear element} \\ 0.5 & \text{for quadratic element} \end{array}\right\} & \text{in 3D} \end{cases} \tag{8.260}$$

The isotropic SC dispersion coefficient D_{sc} (8.257) or (8.258) is added to the hydrodynamic dispersion tensor $\boldsymbol{D}$ (8.242). It yields

$$\boldsymbol{D} = (\varepsilon s D + D_{\mathrm{sc}})\boldsymbol{\delta} + \boldsymbol{D}_{\mathrm{mech}} \tag{8.261}$$

The SC dispersion coefficient $D_{\mathrm{sc}} = D_{\mathrm{sc}}(\phi)$ is nonlinear due to the solution dependency and an appropriate numerical treatment is required. In the practical implementation the SC method is not used in combination with the SU stabilization, i.e., SC stabilizes completely the solution via the isotropic SC dispersion coefficient D_{sc}.

8.14.5 Petrov-Galerkin Least Square (PGLS) Finite Element Method

The Petrov-Galerkin least square (PGLS) FEM represents an alternative stabilization technique to solve transient ADE in the convective form [276]. Its special feature is in introducing a *symmetric* stabilization term. In contrast to the PGFEM, SU and SC upwind methods as described in Sects. 8.14.2–8.14.4, PGLS leads to symmetric matrix systems and possesses built-in streamline-like upwind characteristics. The PGLS symmetrization is superior to symmetric-matrix time integration schemes, where the advective term is treated only explicitly so as done by Leismann and Frind [338]. The effect of PGLS has similarities to the SU upwinding, where an anisotropic (streamline-oriented) balancing dissipation (dispersion) is added to the physical longitudinal dispersion parameter. However, in the PGLS method the artificial dispersion (diffusion) is directly derived from the least-square (LS) finite element concept and requires no 'free' upwind parameter such as α of the preceding upwind methods.

As indicated in Table 8.1 the LS minimization by PGLS has to be done with respect to the nodal values of the state variable(s). Due to the square operations in the inner products of the governing PDE, higher order derivatives remain, which usually require higher order basis functions, i.e., a C_0 continuity (cf. Sect. 8.7) in the interpolation functions is no more sufficient, unless the LS operation is only restricted to first-order terms while the higher order terms are treated in the standard Galerkin-based manner via an operator splitting approach. Basic work was given by Nguyen and Reynen [385] and further developments can be found in [319, 559], among others. König [319] used an operator splitting method in a two-pass strategy, where the separate equations for the diffusive and the advective parts are solved successively. On the other hand, Wendland [559] improved the operator splitting technique by introducing a suited one-pass approach termed as *symmetric streamline stabilization*, where the diffusive and advective parts are reassembled in one symmetric matrix system.

8.14.5.1 Operator Splitting

The basic ADE in the convective form (8.5)

$$\acute{\mathcal{R}}\dot{\phi} + \boldsymbol{q} \cdot \nabla\phi - \nabla \cdot (\boldsymbol{D} \cdot \nabla\phi) + (\vartheta + Q)\phi - H - Q_{\phi w} = 0 \tag{8.262}$$

can be written in an operator-split formulation

$$\acute{\mathcal{R}}\dot{\phi} + (\mathcal{L}^d + \mathcal{L}^a)\phi = H + Q_{\phi w} \tag{8.263}$$

with

$$\begin{aligned} \mathcal{L}^d &= -\nabla \cdot (\boldsymbol{D} \cdot \nabla) + (\vartheta + Q) \\ \mathcal{L}^a &= \boldsymbol{q} \cdot \nabla \end{aligned} \tag{8.264}$$

where $\mathcal{L}^d$ is a diffusion differential operator and $\mathcal{L}^a$ is an advection differential operator. We can also split the solution ϕ into the diffusive and the advective part such that

$$\phi = \phi^d + \phi^a \tag{8.265}$$

Then, we transform (8.263) into two separate equations: first, the diffusive PDE

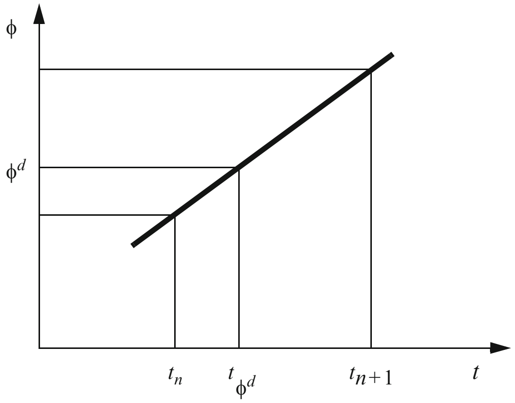


Fig. 8.29 Temporally discrete interpolation of the intermediate diffusive solution ϕ^d

$$\acute{\mathcal{R}}\dot{\phi}^d + \mathcal{L}^d\phi = H + Q_{\phi w} \tag{8.266}$$

and, second, the purely advective (hyperbolic) PDE

$$\acute{\mathcal{R}}(\dot{\phi} - \dot{\phi}^d) + \mathcal{L}^a\phi = 0 \tag{8.267}$$

Summing (8.266) and (8.267) we realize the original ADE (8.263).

The idea of the operator splitting technique is in approximating the diffusive PDE (8.266) and advective PDE (8.267) in a separate manner. After completion the total discrete ADE is obtained by assembling the diffusive and advective parts. In doing so, we consider the variables $\phi^d(t)$ and $\phi^a(t)$ in the time interval (t_n, t_{n+1}) and assume at the beginning of the interval the following IC's for the diffusive variable $\phi_n^d = \phi^d(t_n)$ and for the advective variable $\phi_n^a = \phi^a(t_n)$:

$$\phi_n^d = \phi_n, \quad \phi_n^a = 0, \qquad \dot{\phi}_n^a = 0 \tag{8.268}$$

It is to be noted that the diffusive solution ϕ^d can be considered as an intermediate solution which represents a temporally discrete interpolation between the previous and the new time plane as evidenced in Fig. 8.29.

8.14.5.2 Approximation of the Diffusive Part

In the context of FEM, the two variables ϕ and ϕ^d are replaced by a continuous approximation (8.16) that assumes the separability of space and time, thus

$$\begin{aligned} \phi(\boldsymbol{x}, t) &\approx \textstyle\sum_i N_i(\boldsymbol{x})\,\phi_i(t) \\ \phi^d(\boldsymbol{x}, t) &\approx \textstyle\sum_i N_i(\boldsymbol{x})\,\phi_i^d(t) \end{aligned} \tag{8.269}$$

where the subscript $i = 1, \ldots, N_{\mathrm{P}}$ denotes the global nodal indices. The Galerkin weak statement of (8.266)

$$\mathrm{GWS} = \int_{\Omega} N_i \Big(\acute{\mathcal{R}} \dot{\phi}^d + \mathcal{L}^d \phi - H - Q_{\phi w} \Big) d\Omega = 0 \tag{8.270}$$

leads after inserting the semidiscrete basis functions (8.269) to the following global matrix system (cf. Sect. 8.9)

$$\boldsymbol{O} \cdot \dot{\boldsymbol{\phi}}^d + \boldsymbol{K}^d \cdot \boldsymbol{\phi} - \boldsymbol{F} = \boldsymbol{0} \tag{8.271}$$

with

$$\boldsymbol{K}^d = \boldsymbol{C} + \boldsymbol{R} + \boldsymbol{B} \tag{8.272}$$

where the matrices $\boldsymbol{O}$, $\boldsymbol{C}$, $\boldsymbol{R}$, $\boldsymbol{B}$ and the vector $\boldsymbol{F}$ are given in (8.103)–(8.105) referred to the convective form. By using the methods of time integration introduced above in Sect. 8.13 and invoking (8.265) and (8.268) with

$$\boldsymbol{\phi}(t_n + \theta \Delta t_n) = \theta(\boldsymbol{\phi}^d_{n+1} + \boldsymbol{\phi}^a_{n+1}) + (1-\theta)\boldsymbol{\phi}_n \approx \theta \boldsymbol{\phi}^d_{n+1} + (1-\theta)\boldsymbol{\phi}_n \tag{8.273}$$

the following matrix systems of the intermediate (diffusive) part result

$$\big(\tfrac{\boldsymbol{O}}{\Delta t_n} + \boldsymbol{K}^d \theta\big) \cdot \boldsymbol{\phi}^d_{n+1} = \big[\tfrac{\boldsymbol{O}}{\Delta t_n} - \boldsymbol{K}^d (1-\theta)\big] \cdot \boldsymbol{\phi}_n + \big(\boldsymbol{F}_{n+1}\theta + \boldsymbol{F}_n(1-\theta)\big) \tag{8.274}$$

for the θ−family of time stepping methods (cf. Sect. 8.13.4) and

$$\big(\tfrac{\boldsymbol{O}}{\theta \Delta t_n} + \boldsymbol{K}^d\big) \cdot \boldsymbol{\phi}^d_{n+1} = \boldsymbol{O} \cdot \big[\tfrac{1}{\theta \Delta t_n} \boldsymbol{\phi}_n + \big(\tfrac{1}{\theta} - 1\big) \dot{\boldsymbol{\phi}}_n\big] + \boldsymbol{F}_{n+1} \tag{8.275}$$

for the predictor-corrector methods (cf. Sect. 8.13.5), where the weighting coefficient $\frac{1}{2} \leq \theta \leq 1$ identifies the different time integration methods.

8.14.5.3 Approximation of the Advective Part

The residual of the advective part (8.267) in form of

$$R = \acute{\mathcal{R}}(\dot{\phi} - \dot{\phi}^d) + \boldsymbol{q} \cdot \nabla \phi \tag{8.276}$$

will be treated by the LS method

$$\frac{\partial}{\partial \phi_i} \int_{\Omega} \tfrac{1}{2} R^2 d\Omega = 0 \tag{8.277}$$

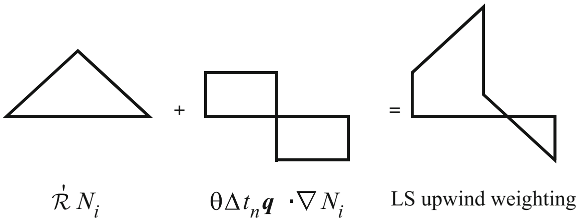

+
=
$\acute{\mathcal{R}} N_i$
$\theta\Delta t_n \boldsymbol{q} \cdot\nabla N_i$
LS upwind weighting

Fig. 8.30 LS weighting function of the operator splitting

which is equivalent to

$$\int_{\Omega} w_i R d\Omega = 0 \tag{8.278}$$

with the nodal test function w_i given by

$$w_i = \frac{\partial R}{\partial \phi_i} \quad (i = 1, \ldots, N_{\mathrm{P}}) \tag{8.279}$$

or with (8.269) in (8.276)

$$w_i = \frac{\partial N_i}{\partial t} + \boldsymbol{q} \cdot \nabla N_i \tag{8.280}$$

Since $N_i = N_i(\boldsymbol{x})$ is not a function of time, the residual (8.276) is to be expressed in its temporally discrete form, such that

$$\begin{aligned} R = \acute{\mathcal{R}}\Big[\sum_i N_i(\phi_{i,n+1} - \phi^d_{i,n+1}) - \sum_i N_i(\phi_{i,n} - \phi^d_{i,n})\Big] + \\ \Delta t_n \boldsymbol{q} \cdot \nabla\Big[\theta \sum_i N_i \phi_{i,n+1} + (1-\theta) \sum_i N_i \phi_{i,n}\Big] \end{aligned} \tag{8.281}$$

which yields

$$w_i = \frac{\partial R}{\partial \phi_{i,n+1}} = \acute{\mathcal{R}} N_i + \theta \Delta t_n \boldsymbol{q} \cdot \nabla N_i \tag{8.282}$$

Then, the LS weak statement (8.278) results

$$\mathrm{LSWS} = \int_{\Omega} \big(\acute{\mathcal{R}} N_i + \theta \Delta t_n \boldsymbol{q} \cdot \nabla N_i\big)\big[\acute{\mathcal{R}}(\dot{\phi} - \dot{\phi}^d) + \boldsymbol{q} \cdot \nabla \phi\big] d\Omega = 0 \tag{8.283}$$

where the residual is weighted by the LS test function (8.282) consisting of two parts as displayed in Fig. 8.30.

The LS weak statement (8.283) leads to the following semidiscrete matrix system

$$(\boldsymbol{O} + \theta \Delta t_n \boldsymbol{V}) \cdot \dot{\boldsymbol{\phi}} + (\boldsymbol{A} + \theta \boldsymbol{T}) \cdot \boldsymbol{\phi} = (\boldsymbol{O} + \theta \Delta t_n \boldsymbol{V}) \cdot \dot{\boldsymbol{\phi}}^d \tag{8.284}$$

with (cf. (8.103))

$$\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \Big(\sum_I \sum_J O^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
\boldsymbol{A} = A_{ij} &= \sum_e \Big(\sum_I \sum_J A^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
\boldsymbol{V} = V_{ij} &= \sum_e \Big(\sum_I \sum_J V^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\
\boldsymbol{T} = T_{ij} &= \sum_e \Big(\sum_I \sum_J T^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big)
\end{aligned} \tag{8.285}$$

and the element matrices

$$\begin{aligned}
O^e_{IJ} &= \int_{\Omega^e} \acute{\mathcal{R}}^e \, N^e_I N^e_J d\Omega^e \\
A^e_{IJ} &= \int_{\Omega^e} N^e_I (\boldsymbol{q}^e \cdot \nabla N^e_J) d\Omega^e \\
V^e_{IJ} &= \int_{\Omega^e} (\boldsymbol{q}^e \cdot \nabla N^e_I) N^e_J d\Omega^e \\
T^e_{IJ} &= \int_{\Omega^e} \Delta t_n \tfrac{1}{\acute{\mathcal{R}}} (\boldsymbol{q}^e \cdot \nabla N^e_I)(\boldsymbol{q}^e \cdot \nabla N^e_J) d\Omega^e
\end{aligned} \tag{8.286}$$

where $\boldsymbol{\Delta}^e$ is the Boolean matrix defined in (8.82). The time discretization of (8.284) for the $\theta-$family of time stepping methods (cf. Sect. 8.13.4) results

$$\begin{aligned}
(\boldsymbol{O} + \theta \Delta t_n \boldsymbol{V}) \cdot \big(\tfrac{\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}_n}{\Delta t_n} \big) + \theta (\boldsymbol{A} + \theta \boldsymbol{T}) \cdot \boldsymbol{\phi}_{n+1} + (1-\theta)(\boldsymbol{A} + \theta \boldsymbol{T}) \cdot \boldsymbol{\phi}_n \\
= (\boldsymbol{O} + \theta \Delta t_n \boldsymbol{V}) \cdot \big(\tfrac{\boldsymbol{\phi}^d_{n+1} - \boldsymbol{\phi}_n}{\Delta t_n} \big)
\end{aligned} \tag{8.287}$$

and finally

$$\big[\tfrac{\boldsymbol{O}}{\Delta t_n} + \theta \big(\boldsymbol{V} + \boldsymbol{A} + \theta \boldsymbol{T} \big) \big] \cdot \boldsymbol{\phi}_{n+1} = \big(\tfrac{\boldsymbol{O}}{\Delta t_n} + \theta \boldsymbol{V} \big) \cdot \boldsymbol{\phi}^d_{n+1} - (1-\theta) \big(\boldsymbol{A} + \theta \boldsymbol{T} \big) \cdot \boldsymbol{\phi}_n \tag{8.288}$$

Regarding the predictor-corrector strategy based on the BE and TR schemes, if taking

$$\begin{aligned}
\dot{\boldsymbol{\phi}}_{n+1} &= \tfrac{1}{\theta \Delta t_n} \big(\boldsymbol{\phi}_{n+1} - \boldsymbol{\phi}_n \big) - (\tfrac{1}{\theta} - 1) \dot{\boldsymbol{\phi}}_n \\
\dot{\boldsymbol{\phi}}^d_{n+1} &= \tfrac{1}{\theta \Delta t_n} \big(\boldsymbol{\phi}^d_{n+1} - \boldsymbol{\phi}_n \big) - (\tfrac{1}{\theta} - 1) \dot{\boldsymbol{\phi}}_n
\end{aligned} \tag{8.289}$$

and using (8.284), the following matrix system for the predictor-corrector schemes is obtained

$$\big(\tfrac{\boldsymbol{O}}{\theta \Delta t_n} + \boldsymbol{V} + \boldsymbol{A} + \theta \boldsymbol{T} \big) \cdot \boldsymbol{\phi}_{n+1} = \big(\tfrac{\boldsymbol{O}}{\theta \Delta t_n} + \boldsymbol{V} \big) \cdot \boldsymbol{\phi}^d_{n+1} \tag{8.290}$$

8.14.5.4 Assembly of the Diffusive and Advective Parts

To obtain the matrix system for the complete ADE (8.262) the diffusive and advective parts have to be added. For the $\theta-$family of time stepping methods the summation of (8.274) and (8.288) yields

$$\begin{aligned}\Big[\tfrac{\boldsymbol{O}}{\Delta t_n} + \theta\big(\boldsymbol{V} + \boldsymbol{A} + \theta\boldsymbol{T}\big)\Big]\cdot\boldsymbol{\phi}_{n+1} + \big(\tfrac{\boldsymbol{O}}{\Delta t_n} + \theta\boldsymbol{K}^d\big)\cdot\boldsymbol{\phi}^d_{n+1} = \big(\tfrac{\boldsymbol{O}}{\Delta t_n} + \theta\boldsymbol{V}\big)\cdot\boldsymbol{\phi}^d_{n+1} \\ -(1-\theta)\big(\boldsymbol{A} + \theta\boldsymbol{T}\big)\cdot\boldsymbol{\phi}_n + \Big[\tfrac{\boldsymbol{O}}{\Delta t_n} - (1-\theta)\boldsymbol{K}^d\Big]\cdot\boldsymbol{\phi}_n + \big(\boldsymbol{F}_{n+1}\theta + \boldsymbol{F}_n(1-\theta)\big)\end{aligned} \tag{8.291}$$

The term $\frac{\boldsymbol{O}}{\Delta t_n}\cdot\boldsymbol{\phi}^d_{n+1}$ can be eliminated from (8.291). The remaining terms correlating with the intermediate solution $\boldsymbol{\phi}^d_{n+1}$ will be transformed in the following way [559]: All terms related to $\boldsymbol{\phi}^d_{n+1}$ on the LHS of (8.291) are replaced by $\boldsymbol{\phi}_{n+1}$, while such terms on the RHS of (8.291) are substituted by $\boldsymbol{\phi}_n$. In doing so, the following matrix system results

$$\begin{aligned}\Big[\tfrac{\boldsymbol{O}}{\Delta t_n} + \theta\big(\boldsymbol{K}^d + \boldsymbol{V} + \boldsymbol{A} + \theta\boldsymbol{T}\big)\Big]\cdot\boldsymbol{\phi}_{n+1} = \\ \Big[\tfrac{\boldsymbol{O}}{\Delta t_n} - (1-\theta)(\boldsymbol{K}^d + \boldsymbol{A} + \theta\boldsymbol{T}) + \theta\boldsymbol{V}\Big]\cdot\boldsymbol{\phi}_n + \big(\boldsymbol{F}_{n+1}\theta + \boldsymbol{F}_n(1-\theta)\big)\end{aligned} \tag{8.292}$$

Analogously, for the BE and TR predictor-corrector schemes we add (8.275) and (8.290)

$$\begin{aligned}\big(\tfrac{\boldsymbol{O}}{\theta\Delta t_n} + \boldsymbol{V} + \boldsymbol{A} + \theta\boldsymbol{T}\big)\cdot\boldsymbol{\phi}_{n+1} + \big(\tfrac{\boldsymbol{O}}{\theta\Delta t_n} + \boldsymbol{K}^d\big)\cdot\boldsymbol{\phi}^d_{n+1} = \\ \big(\tfrac{\boldsymbol{O}}{\theta\Delta t_n} + \boldsymbol{V}\big)\cdot\boldsymbol{\phi}^d_{n+1} + \boldsymbol{O}\cdot\Big[\tfrac{1}{\theta\Delta t_n}\boldsymbol{\phi}_n + \big(\tfrac{1}{\theta} - 1\big)\dot{\boldsymbol{\phi}}_n\Big] + \boldsymbol{F}_{n+1}\end{aligned} \tag{8.293}$$

which gives

$$\big(\tfrac{\boldsymbol{O}}{\theta\Delta t_n} + \boldsymbol{K}^d + \boldsymbol{V} + \boldsymbol{A} + \theta\boldsymbol{T}\big)\cdot\boldsymbol{\phi}_{n+1} = \boldsymbol{V}\cdot\boldsymbol{\phi}_n + \boldsymbol{O}\cdot\Big[\tfrac{1}{\theta\Delta t_n}\boldsymbol{\phi}_n + \big(\tfrac{1}{\theta} - 1\big)\dot{\boldsymbol{\phi}}_n\Big] + \boldsymbol{F}_{n+1} \tag{8.294}$$

The final matrix systems (8.292) and (8.294) for the $\theta-$family of time stepping methods and the BE and TR predictor-corrector schemes, respectively, are symmetric and positive definite. It results from the fact that the advective matrices $\boldsymbol{V}$ and $\boldsymbol{A}$ form a symmetric contribution as the sum $(\boldsymbol{V} + \boldsymbol{A})$ because $\boldsymbol{A} = \boldsymbol{V}^T$ is the transpose as easily seen from (8.286). This is only attainable for the ADE convective form (8.5) or (8.262). Unlikely, the ADE divergence form (8.3) is rather inappropriate for the PGLS method.[15]

[15] The ADE divergence form (8.3) contains a divergence expression of the advection term in form of $\nabla\cdot(\boldsymbol{q}\phi)$. For the split advective part (8.267) the advective operator $\mathcal{L}^a$ would be

$$\mathcal{L}^a = \boldsymbol{q}\cdot\nabla + (\nabla\cdot\boldsymbol{q})$$

The symmetric term $\boldsymbol{T}$ appearing in (8.292) and (8.294) can be interpreted as an additional term of artificial diffusion. This naturally results from the LS weighting procedure (8.277). In comparison to the SU method (see Sect. 8.14.3) where an anisotropic balancing dissipation tensor $\boldsymbol{D}_{\text{num}} = \beta_{\text{num}} \frac{\boldsymbol{q} \otimes \boldsymbol{q}}{\|\boldsymbol{q}\|}$ (8.243) with β_{num} (8.245) as a function of the element length h^e and the upwind parameter α is added, it is obvious that the LS damping matrix $\boldsymbol{T}$ (8.285), (8.286) is identical to $\boldsymbol{D}_{\text{num}}$ except for the parameter β_{num}, which becomes for the PGLS method

$$\beta_{\text{num}} = \frac{\Delta t_n \|\boldsymbol{q}^e\|}{\acute{\mathcal{R}}} = Cr\, h^e \tag{8.295}$$

where Cr is the Courant number defined in (8.236). It reveals that the PGLS upwinding is quite similar to a SU method, where the damping is performed in the longitudinal direction of flow via the added damping parameter β_{num}. While in the SU method β_{num} is a function on the element length h^e and the free upwind parameter $0 \leq \alpha \leq 1$, in the PGLS the damping parameter β_{num} is dependent on the time step size Δt_n and the quotient $\|\boldsymbol{q}^e\|/\acute{\mathcal{R}}$. Hence, PGLS is recognized as a built-in streamline-like upwind strategy, however, without any free upwind parameter. Comparing β_{num} for the PGLS of (8.295) with the SU method of (8.245) it is apparent that the Courant number should be $Cr \leq 0.5$ for the PGLS (at linear elements) to avoid an overdamping larger than in the SU method.

8.14.6 An Illustrative Example

To demonstrate the impact of the different finite element schemes introduced above on stability and accuracy we consider a representative example of an advection-

Then, the LS weak statement of the advective part is

$$\text{LSWS} = \int_{\Omega} \left[\acute{\mathcal{R}} N_i + \theta \Delta t_n \nabla \cdot (\boldsymbol{q} N_i)\right]\left[\acute{\mathcal{R}}(\dot{\phi} - \dot{\phi}^d) + \nabla \cdot (\boldsymbol{q}\phi)\right] d\Omega = 0$$

which leads to a matrix system equivalent to (8.284), but having different element matrices

$$\begin{aligned}
\boldsymbol{A}^e &= \int_{\Omega^e} N_I^e[(\boldsymbol{q}^e \cdot \nabla N_J^e) + (\nabla \cdot \boldsymbol{q}^e) N_J^e] d\Omega^e \\
\boldsymbol{V}^e &= \int_{\Omega^e} [(\boldsymbol{q}^e \cdot \nabla N_I^e) + (\nabla \cdot \boldsymbol{q}^e) N_I^e] N_J^e d\Omega^e \\
\boldsymbol{T}^e &= \int_{\Omega^e} \Delta t_n \tfrac{1}{\acute{\mathcal{R}}} [\nabla \cdot (\boldsymbol{q}^e N_I^e)][\nabla \cdot (\boldsymbol{q}^e N_J^e)] d\Omega^e
\end{aligned}$$

While the symmetry of the matrix system is still maintained since $\boldsymbol{A}^e = \boldsymbol{V}^{eT}$, the divergence expressions $(\nabla \cdot \boldsymbol{q}^e)$ appearing in $\boldsymbol{A}^e$, $\boldsymbol{V}^e$ and $\boldsymbol{T}^e$ can cause difficulties if the flow is not selenoidal (i.e., not divergence-free: $\nabla \cdot \boldsymbol{q} \neq 0$) at the presence of storage and sources/sinks. This makes the LS technique rather inappropriate for the ADE divergence form.

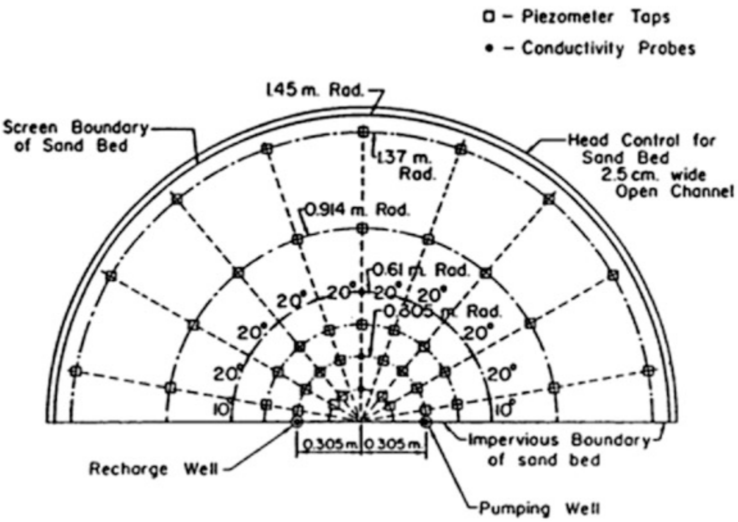


Fig. 8.31 Plane view of Hoopes and Harlemann's sand-filled semi-cylinder [257]

dominated solute transport on a nonuniform flow field to which analytical and experimental results are available. It is known as the Hoopes and Harlemann's two-well problem [257, 470]. Hoopes and Harlemann [257] performed a lab-scale experiment in a semi-cylinder filled with sand as shown in Fig. 8.31. 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 2D horizontal confined aquifer which is homogeneous and isotropic. The nonuniform flow between the well doublet at a distance, $2d = 0.61$ m, is isothermal and in a steady state. The solute transport is only affected by advection and dispersion. Comparisons of the analytical result with experiments and various numerical solution schemes have already been performed elsewhere [136, 257, 282, 470]. The Hoopes and Harlemann's problem is now used to compare the different numerical schemes with the analytical (exact) results.

One obtains the 2D analytical solution in terms of the velocity potential Φ and the streamline function Ψ (cf. Sect. 2.1.11). They are related to the original x_1, x_2–coordinates via the conformal transformation

$$\Phi + i\Psi = \mathrm{Ln}(z+d)/(z-d), \quad \text{with} \quad z = x_1 + i x_2 \tag{8.296}$$

where Ln is the complex natural logarithm, i corresponds to the imaginary unit and d is the half well spacing. This transformation maps the area of the half circle with radius $r \leq d$ onto a strip of infinite length and width $\pi/2$. The governing transport equation can now be transformed to a 1D equation written in the form

$$\frac{\partial \phi}{\partial t} + v^2 \left(\frac{\partial \phi}{\partial \Phi} + D \frac{\partial^2 \phi}{\partial \Phi^2} \right) = 0 \tag{8.297}$$

where v is the intrinsic velocity and $D = D_o + \beta_L v$ is the dispersion coefficient (D_0 = molecular diffusion, β_L = longitudinal dispersivity). The intrinsic velocity v at a flux rate Q_w of the recharge well is given by

$$v = \frac{Q_w}{2\pi B \acute{\mathcal{R}} d}\big(\cosh(\Phi_D) + \cos(\Psi_D)\big) \tag{8.298}$$

with the dimensionless quantities

$$\begin{aligned} \Phi_D &= \tfrac{1}{2}\ln\Big(\tfrac{(x_1+d)^2+x_2^2}{(x_1-d)^2+x_2^2}\Big) \\ \Psi_D &= \arctan\Big(\tfrac{-2x_2 d}{x_1^2+x_2^2-d^2}\Big) \end{aligned} \tag{8.299}$$

where B is the aquifer thickness. The velocity potential $\Phi(x_1, x_2)$ and the streamfunction $\Psi(x_1, x_2)$ are obtained after multiplication with $Q_w/(2\pi B)$. The IC and BC are

$$\phi(\Phi, \Psi, t_0) = 0, \quad \text{and} \quad \phi(\Phi(-d, 0), \Psi(-d, 0), t) = \phi_w \tag{8.300}$$

The dimensionless solutal quantity at arbitrary time is given by

$$\phi_D = \frac{\phi}{\phi_w} = \tfrac{1}{2}\text{erfc}\Big(\frac{I_D - t_D}{2\sqrt{J_D}}\Big) \tag{8.301}$$

where erfc() is the complementary error function and $t_D = Q_w t/(2\pi B \acute{\mathcal{R}} d^2)$ is the dimensionless time. Owing to the properties of the conformal transformation (8.296), the solution (8.301) can be calculated for given spatial points, where the integrals are

$$I_D = \int_{-\infty}^{\Phi_D} \frac{d\Phi_D}{v_D^2}, \quad \text{and} \quad J_D = \int_{-\infty}^{\Phi_D} \frac{D_D}{v_D^4} d\Phi_D \tag{8.302}$$

in which $v_D = 2\pi B \acute{\mathcal{R}} d v / Q_w$ and $D_D = 2\pi B \acute{\mathcal{R}} D / Q_w$ are likewise dimensionless. The complete analytical solution is given in [257, 470], but, its evaluation is cumbersome.

For the present comparative study the finite element computations are performed on a triangle mesh of the symmetric half of the circular problem as shown in Fig. 8.32. The mesh is consciously chosen relatively coarse and only slightly refined in the vicinity of the recharge and pumping wells, where high velocities are expected. The used model parameters are listed in Table 8.8. The AB/TR predictor-corrector method with automatic time stepping is firstly employed. Hoopes and Harlemann [257] assumed no dispersion across streamlines in their formulation of (8.297). The longitudinal dispersivity $\beta_L = 0.0015\,\text{m}$ is very small and gives rise to a steep front of solute in space and time. Hence, the transport is dominated

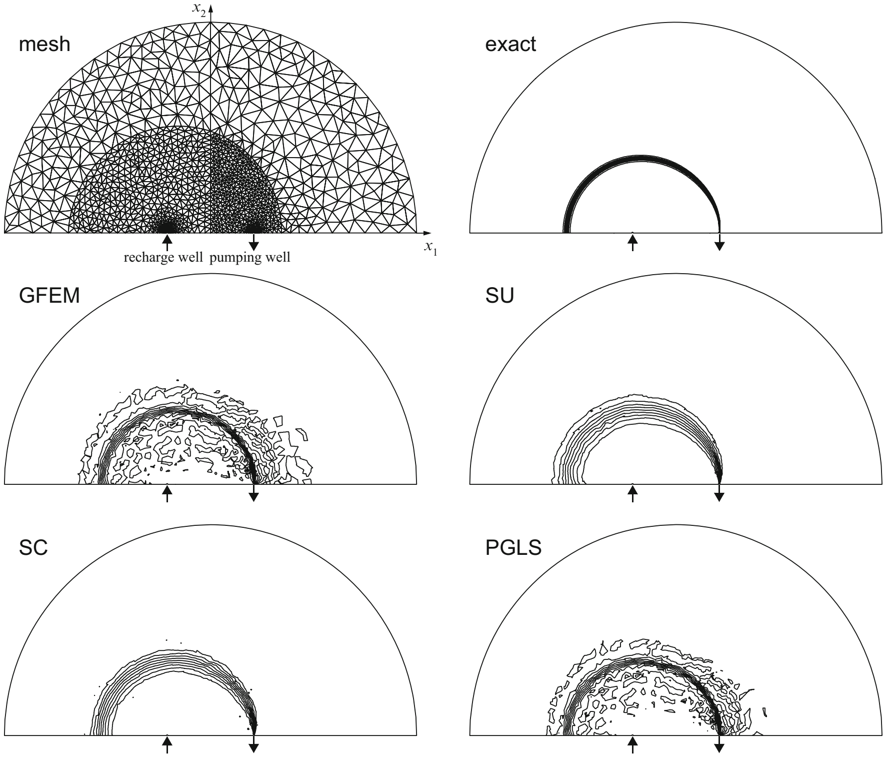


Fig. 8.32 Used 2D mesh consisting of 4,890 triangles with 2,544 nodes and resulting isocontours of solution ϕ at $t = 0.2$ d computed by GFEM, SU, SC and PGLS using adaptive AB/TR time stepping in comparison to the exact distribution

by advection as evidenced in Fig. 8.33 for breakthrough curves of two observation points located at $x_2 = 0.145$ m and $x_2 = 0.305$ m along the symmetry line with $x_1 = 0$ m between the wells. The breakthrough histories obtained for the different finite element schemes are compared to the exact curve. As seen the standard GFEM method provides oscillating numerical solutions, while the SU and SC schemes can completely dampen out the oscillations but introduce in turn spurious numerical dispersion. Figure 8.33 reveals that the PGLS scheme is not able to produce wiggle-free solutions. Since the stabilization in the PGLS is dependent on the actual time step size, cf. the relationship of (8.295), it is obvious that the time steps generated by the adaptive predictor-corrector procedure are unsuitably small to achieve sufficient damping via the LS mechanism.

The simulated solute distributions at $t = 0.2$ d for the GFEM, SU, SC and PGLS schemes by using AB/TR time stepping are depicted in Fig. 8.32 for isocontours of ten solute levels spanning between the maximum and minimum values of the attained numerical results. The exact solution also shown in Fig. 8.32 reveals a sharp circular-shaped front, which can be hardly modeled by the numerical methods on the used coarse mesh. Thus, the GFEM and the PGLS schemes exhibit bounded

Table 8.8 Simulation parameters of the Hoopes and Harlemann's two-well problem

Quantity	Symbol	Value	Unit
Flow			
Well discharge	Q_w	$2.339 \cdot 10^{-6}$	$\mathrm{m}^3\,\mathrm{s}^{-1}$
Flux at recharge well	$\bar{q}_{n_h} = \frac{Q_w}{2\pi R}$	6.4327	$\mathrm{m}^2\,\mathrm{d}^{-1}$
Head at pumping well	h_D	0	m
Isotropic aquifer transmissivity	T	10^{-4}	$\mathrm{m}^2\,\mathrm{s}^{-1}$
Solute transport			
Initial condition	$\phi(\boldsymbol{x}, t_0)$	0	$\mathrm{mg}\,\mathrm{l}^{-1}$
Solute at recharge well	ϕ_w	1	$\mathrm{mg}\,\mathrm{l}^{-1}$
Aquifer thickness	B	0.089	m
Porosity	ε	0.374	
Molecular diffusion	D_o	0	$\mathrm{m}^2\,\mathrm{s}^{-1}$
Longitudinal dispersivity	β_L	0.0015	m
Transverse dispersivity	β_T	0	m
FEM			
Half well spacing	d	0.305	m
Wellbore radius	R	0.005	m
Outer boundary radius	R_Ω	1.45	m
Number of triangular elements	N_E	4,890	
Number of mesh nodes	N_P	2,544	
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (AB/TR method)	ϵ	10^{-4}	1
Simulation time period	t_end	0.2	d

oscillations, which also spread in a distance from the front. On the other hand, the SU scheme results a quite wiggle-free distribution, however, the front is significantly widened due to the numerical dispersion effect. However, it is interesting that the SC scheme, while also providing essentially non-oscillatory solutions, can remarkably reduce the amount of spurious numerical dispersion and gives reasonably better results than the SU method.

To complete the comparison let us also investigate the 1st-order accurate FE/BE time stepping method, which promises a higher stability. Indeed, in this case even the GFEM leads to well-stabilized solutions as displayed in Fig. 8.34, however, we note a remarkable influence of numerical dispersion if the time steps are chosen large (we enforce large time steps by relaxing the RMS error criterion in the FE/BE predictor corrector method according to $\epsilon = 10^{-2}$). To better understand the reason and measure of this effect, in the next section the quantities of numerical dispersion will be estimated for the different methods. Further numerical comparisons in 2D and 3D applications can be found in [136].

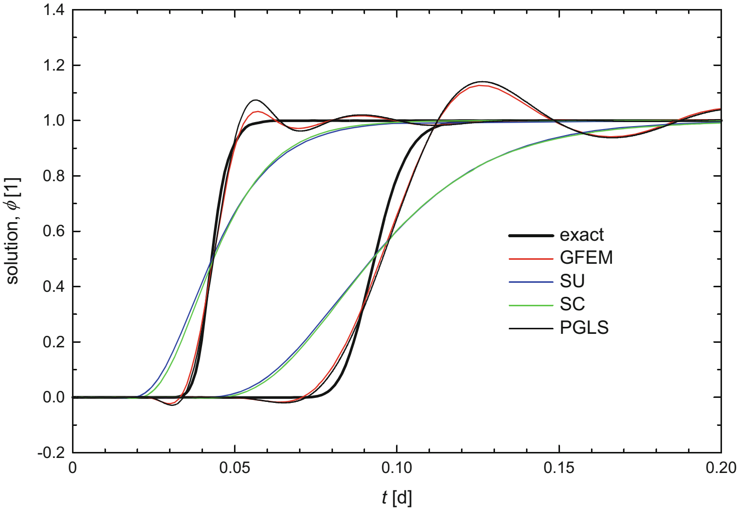


Fig. 8.33 Breakthrough of approximate solutions obtained with the 2nd-order accurate AB/TR time integration method in comparison to the exact history at two points located at $x_2 = 0.145\,\text{m}$ (*left*) and $0.305\,\text{m}$ (*right*) on the symmetric line $x_1 = 0$ between the wells

8.15 Summarized Quantitative Discussion of Error and Stability for the Favorite Schemes

A quantitative error estimate of the spatio-temporally discrete equations can be obtained by evaluating the LTE associated with the temporal and spatial derivatives. For this purpose let us consider the discrete equations (8.223) resulting from a PGFEM approximation of the 1D ADE convective form without sources/sinks and boundary terms for a uniform mesh with linear elements of length h $(= \Delta x)$, CM matrix and constant parameters written as

$$\begin{aligned}
&\frac{\acute{\mathcal{R}}}{6}\Big(\frac{\phi_{i+1,n+1}-\phi_{i+1,n}}{\Delta t_n}\Big)+\frac{2\acute{\mathcal{R}}}{3}\Big(\frac{\phi_{i,n+1}-\phi_{i,n}}{\Delta t_n}\Big)+\frac{\acute{\mathcal{R}}}{6}\Big(\frac{\phi_{i-1,n+1}-\phi_{i-1,n}}{\Delta t_n}\Big)+\\
&\qquad\theta\Big[-\frac{D}{h^2}+(1-\alpha)\frac{q}{2h}\Big]\phi_{i+1,n+1}+(1-\theta)\Big[-\frac{D}{h^2}+(1-\alpha)\frac{q}{2h}\Big]\phi_{i+1,n}+\\
&\qquad\qquad\theta\Big(\frac{2D}{h^2}+\alpha\frac{q}{h}\Big)\phi_{i,n+1}+(1-\theta)\Big(\frac{2D}{h^2}+\alpha\frac{q}{h}\Big)\phi_{i,n}+\\
&\qquad\theta\Big[-\frac{D}{h^2}-(1+\alpha)\frac{q}{2h}\Big]\phi_{i-1,n+1}+(1-\theta)\Big[-\frac{D}{h^2}-(1+\alpha)\frac{q}{2h}\Big]\phi_{i-1,n}=0
\end{aligned}\tag{8.303}$$

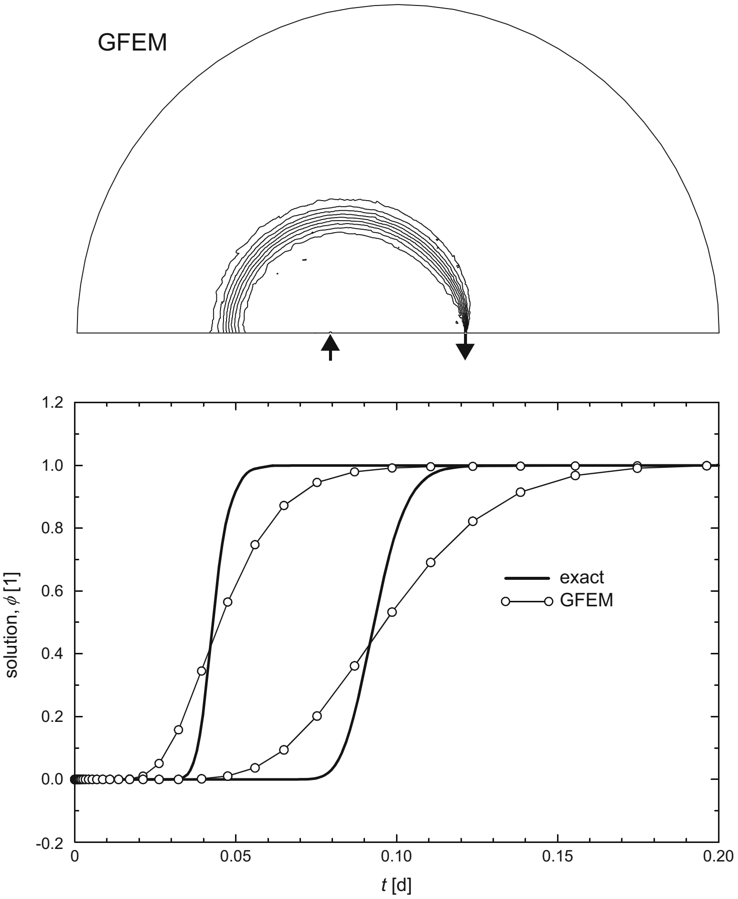


Fig. 8.34 Isocontours of solution at $t = 0.2$ d and breakthrough curves simulated for GFEM by using the 1st-order accurate FE/BE predictor-corrector method with a relaxed RMS error criterion of $\epsilon = 10^{-2}$

A Taylor series expansion in time for $\phi_{i+1,n}$ about the time plane $n+1$ and for $\phi_{i+1,n+1}$ about the time plane n gives, cf. (8.149)

$$\left(\frac{\phi_{i+1,n+1}-\phi_{i+1,n}}{\Delta t_n}\right) = \theta\Big[\dot{\phi}_{i+1} - \frac{\Delta t_n}{2}\ddot{\phi}_{i+1} + \frac{\Delta t_n^2}{6}\dddot{\phi}_{i+1} - \mathcal{O}(\Delta t_n^3)\Big]_{n+1} + (1-\theta)\Big[\dot{\phi}_{i+1} + \frac{\Delta t_n}{2}\ddot{\phi}_{i+1} + \frac{\Delta t_n^2}{6}\dddot{\phi}_{i+1} + \mathcal{O}(\Delta t_n^3)\Big]_{n} \tag{8.304}$$

Similar expressions result for $(\phi_{i,n+1}-\phi_{i,n})/\Delta t_n$ and $(\phi_{i-1,n+1}-\phi_{i-1,n})/\Delta t_n$.

Now, we can also apply a Taylor series expansion in space for $\phi_{i+1,n+1}$ and $\phi_{i-1,n+1}$, respectively, about i at a given time plane $n+1$ to obtain

$$\begin{aligned}\phi_{i+1,n+1} &= \left[\phi_i + h\phi_i' + \frac{h^2}{2}\phi_i'' + \frac{h^3}{6}\phi_i''' + \mathcal{O}(h^4)\right]_{n+1}\\ \phi_{i-1,n+1} &= \left[\phi_i - h\phi_i' + \frac{h^2}{2}\phi_i'' - \frac{h^3}{6}\phi_i''' + \mathcal{O}(h^4)\right]_{n+1}\end{aligned} \tag{8.305}$$

where $'$ denotes differentiation with respect to the 1D space coordinate $\partial/\partial x$. Similar expressions result for ϕ_{i+1} and ϕ_{i-1} at the other time planes n and $n-1$, i.e., $\phi_{i+1,n}$, $\phi_{i-1,n}$, $\phi_{i+1,n-1}$ and $\phi_{i-1,n-1}$, respectively.

To obtain expressions for higher order time derivatives $\ddot{\phi}$ and $\dddot{\phi}$ in terms of spatial derivatives, the governing 1D ADE $\dot{\phi}+q\phi'-D\phi''=0$ may be rewritten in the form

$$\dot{\phi} = -q\phi' + D\phi'' \tag{8.306}$$

Differentiating (8.306) with respect to time, rearranging the differentials and successively substituting again (8.306) in the resulting expression, gives

$$\ddot{\phi} = \frac{\partial}{\partial t}\Big(-q\phi' + D\phi''\Big) = -q\frac{\partial}{\partial x}\dot{\phi} + D\frac{\partial}{\partial x^2}\dot{\phi} = q^2\phi'' - 2qD\phi''' + D^2\phi^{(4)} \tag{8.307}$$

and similarly

$$\dddot{\phi} = -q^3\phi''' + 3q^2D\phi^{(4)} - 3qD^2\phi^{(5)} + D^3\phi^{(6)} \tag{8.308}$$

Now, inserting (8.304) and the related expressions into (8.303), replacing all higher order time derivatives by spatial derivatives via (8.307) and (8.308) as well as substituting with (8.305) all $(i+1)$th and $(i-1)$th terms, we find after some manipulations the following approximate representation of the governing ADE at node i and time plane $n+\theta$ (note that $t_{n+\theta} = \theta t_{n+1} + (1-\theta)t_n$):

$$\begin{aligned}\left(\acute{\mathcal{R}}\dot{\phi} + q\phi' - D\phi''\right)_{i,n+\theta} = &\Big(\alpha\tfrac{qh}{2} + \tfrac{\Delta t_n}{2}q^2\Big)\theta\phi''_{i,n+1} + \Big(\alpha\tfrac{qh}{2} - \tfrac{\Delta t_n}{2}q^2\Big)(1-\theta)\phi''_{i,n}\\ &+\Big(\tfrac{\Delta t_n^2}{6}q^3 - \Delta t_n qD\Big)\theta\phi'''_{i,n+1} + \Big(\tfrac{\Delta t_n^2}{6}q^3 + \Delta t_n qD\Big)(1-\theta)\phi'''_{i,n}\\ &+\Big(\tfrac{h^2}{12}D - \tfrac{\Delta t_n}{2}D^2 + \tfrac{\Delta t_n^2}{2}q^2D + \tfrac{\Delta t_n^3}{24}q^4 - \alpha\tfrac{q}{24}h^3\Big)\theta\phi^{(4)}_{i,n+1}\\ &+\Big(\tfrac{h^2}{12}D + \tfrac{\Delta t_n}{2}D^2 + \tfrac{\Delta t_n^2}{2}q^2D + \tfrac{\Delta t_n^3}{24}q^4 - \alpha\tfrac{q}{24}h^3\Big)(1-\theta)\phi^{(4)}_{i,n}\\ &+\text{HOT}\end{aligned} \tag{8.309}$$

or simply

$$\acute{\mathcal{R}}\dot{\phi} + q\phi' = (D + D_{\text{num}})\phi'' + \mathcal{O}\big(h^2, (2\theta-1)\Delta t_n, \Delta t_n^2\big) \tag{8.310}$$

where the *numerical dispersion* coefficient D_{num} associated with the second spatial derivatives on the RHS of (8.309) appears

$$D_{\text{num}} = \alpha \frac{qh}{2} + \Delta t_n q^2 (\theta - \tfrac{1}{2}) \tag{8.311}$$

The RHS of (8.309) encompasses the total truncation error of the spatio-temporal discretization, which has no physical basis. The coefficient D_{num} of the *spurious* (unphysical) numerical dispersion covers the leading terms of the truncation errors, which are of 1st order in space and time.

In generalization of (8.310) we can find a semidiscrete representation of the governing ADE convective form in multidimensions as

$$\acute{\mathcal{R}} \frac{\partial \phi}{\partial t} + \boldsymbol{q} \cdot \nabla \phi - \nabla \cdot [(\boldsymbol{D} + \boldsymbol{D}_{\text{num}}) \cdot \nabla \phi] + (\vartheta + Q)\phi - H - Q_{\phi w} = \mathcal{O}\big(h^{e2}, (2\theta - 1)\Delta t_n, \Delta t_n^2\big) \tag{8.312}$$

with the tensor of numerical dispersion

$$\boldsymbol{D}_{\text{num}} = D_{\text{num}} \boldsymbol{\delta} + \beta_{\text{num}} \tfrac{\boldsymbol{q} \otimes \boldsymbol{q}}{\|\boldsymbol{q}\|} \tag{8.313}$$

where β_{num} is the streamline-oriented coefficient of numerical dispersion while the scalar numerical dispersion D_{num} consists of the spatial part $D_{\text{num}}^{\text{space}}$ and the temporal part $D_{\text{num}}^{\text{time}}$, viz.,

$$D_{\text{num}} = D_{\text{num}}^{\text{space}} + D_{\text{num}}^{\text{time}} \tag{8.314}$$

with

$$\begin{aligned} D_{\text{num}}^{\text{space}} &\sim \text{different for GFEM, SU, FU, SC, PGLS} \\ D_{\text{num}}^{\text{time}} &= \Delta t_n q^{e2} (\theta - \tfrac{1}{2}) \end{aligned} \tag{8.315}$$

The coefficients of numerical dispersion, the orders of accuracy and the stability restrictions for the different favorite schemes discussed above in Sects. 8.13 and 8.14 are summarize in Table 8.9, where the parameters are evaluated at element level. The standard Galerkin FEM with the TR (Crank-Nicolson) time stepping is recognized as the most accurate method which is 2nd-order accurate in space and time $\mathcal{O}(h^{e2}, \Delta t_n^2)$ without numerical dispersion, however, it is only conditionally stable. Most restrictive and crucial for the GFEM is the $Pg < 1$ condition, unless oscillatory solutions can be produced. The diffusion limit $Cr < Pg$ is commonly not important for the TR (Crank-Nicolson) scheme, while the advection-diffusion limit $PgCr < 1$ can be more serious for dominant advection if using the TR time stepping. In Table 8.9 the accuracy of the schemes decreases from top to down in favor of increasing stability. However, the higher stability is paid by an increased amount of spurious numerical dispersion which can significantly exceed the physical dispersion/diffusion $\boldsymbol{D}^e$ if the mesh is coarse and/or the time steps are large. Here, the FU scheme with fully implicit time stepping of 1st-order accuracy $\mathcal{O}(h^e, \Delta t_n)$, while unconditionally stable, tends to produce very overdiffusive results if h^e and/or Δt_n are large. Compromises between stability and accuracy are possible

Table 8.9 Estimated accuracy and stability restrictions of the favorite schemes using linear finite elements and semi-implicit or implicit ($\mathcal{A}-$stable) time integration in solving the ADE (8.5) or (8.3) at presence of advection $\boldsymbol{q} \neq \boldsymbol{0}$. Note that for diffusion/conduction problems ($\boldsymbol{q} \equiv \boldsymbol{0}$) only the standard GFEM is applied possessing no numerical dispersion and no stability restrictions, except the diffusion limit $Cr < Pg$, ($\Delta t_n < \acute{\mathcal{R}} h^{e2}/(2D^e)$), for the 2nd-order accurate TR time integration method

	Time	Accuracy				Stability[a]		
	integra-	Numerical dispersion[b]				Temporal	Spatial	
Scheme	tion[c]	D^e_{num}	β^e_{num}	Order		limits[d]	limits	
GFEM[f]	TR	–	–	$\mathcal{O}(h^{e2}, \Delta t_n^2)$	high	$Cr < Pg < \frac{1}{Cr}$	$Pg < 1$	low
	BE	$\frac{\Delta t_n}{2} q^{e2}$	–	$\mathcal{O}(h^{e2}, \Delta t_n)$	↑	–	$Pg < 1$	↓
PGLS[g]	TR	–	$Cr\, h^e$	$\mathcal{O}(h^{e2}, \Delta t_n)$		$Cr < Pg < \frac{1}{Cr}$	–	
	BE	$\frac{\Delta t_n}{2} q^{e2}$	$Cr\, h^e$	$\mathcal{O}(h^{e2}, \Delta t_n)$		–	–	
SU[h]	TR	–	$\frac{h^e}{2}$	$\mathcal{O}(h^e, \Delta t_n^2)$		$Cr < Pg < \frac{1}{Cr}$	–	
	BE	$\frac{\Delta t_n}{2} q^{e2}$	$\frac{h^e}{2}$	$\mathcal{O}(h^e, \Delta t_n)$		–	–	
SC[i]	TR	$\frac{1}{2}\alpha_c h^e \Vert\boldsymbol{q}_{\Vert}\Vert$	–	$\mathcal{O}(h^e, \Delta t_n^2)$		$Cr < Pg < \frac{1}{Cr}$	–	
	BE	$\frac{1}{2}\alpha_c h^e \Vert\boldsymbol{q}_{\Vert}\Vert + \frac{\Delta t_n}{2} q^{e2}$	–	$\mathcal{O}(h^e, \Delta t_n)$		–	–	
FU[j]	TR	$\frac{h^e}{2}\Vert\boldsymbol{q}^e\Vert$	–	$\mathcal{O}(h^e, \Delta t_n^2)$		$Cr < Pg < \frac{1}{Cr}$	–	
	BE	$\frac{1}{2}(h^e + \Delta t_n \Vert\boldsymbol{q}^e\Vert)\Vert\boldsymbol{q}^e\Vert$	–	$\mathcal{O}(h^e, \Delta t_n)$	low	–	–	high

[a] Necessary but not always sufficient to ensure boundedness and prevent oscillations

[b] Expressed in the element tensor $\boldsymbol{D}^e_{\text{num}} = D^e_{\text{num}}\boldsymbol{\delta} + \beta^e_{\text{num}}\frac{\boldsymbol{q}^e \otimes \boldsymbol{q}^e}{\Vert\boldsymbol{q}^e\Vert}$ of numerical dispersion (8.313)

[c] $\theta-$family and corrector methods: $\theta = \frac{1}{2}$, TR (Crank-Nicolson); $\theta = 1$, BE (fully implicit)

[d] Conditionally stable, $Cr = \frac{\Vert\boldsymbol{q}^e\Vert \Delta t_n}{\acute{\mathcal{R}} h^e}$, $Pg = \frac{\Vert\boldsymbol{q}^e\Vert h^e}{2\Vert\boldsymbol{D}^e\Vert}$ ($\boldsymbol{D}^e =$ physical dispersion, h^e by (8.239))

[f] Standard Galerkin without any upwinding, $\alpha = 0$

[g] Least square strategy suitable for ADE convective form

[h] Streamline upwinding used with $\alpha = 1$, (8.245)

[i] Shock capturing with projected flux $\Vert\boldsymbol{q}^e_{\Vert}\Vert = \frac{\vert\boldsymbol{q}^e \cdot \nabla\phi^e\vert}{\Vert\nabla\phi^e\Vert}$ and upwind parameter α_c, (8.259)

[j] Full upwinding equivalent to (8.251) with $\beta^e_{\text{num}} = \frac{h^e}{2}$

by resorting to the PGLS, SU or SC schemes, where always the 1st-order accurate BE time stepping provides a higher stability in contrast to the 2nd-order accurate TR time stepping scheme.

In predictor-corrector time stepping only the corrector solutions are important for the stability analysis, while the explicit predictors provide prolongated solutions, which are primarily used to estimate the accuracy in comparison to the corrector solutions needed in the adaptive time stepping control. However, the accuracy of the predictor and corrector must be consistent, so that FE and BE are both 1st-order accurate in time as well as the AB and TR are both 2nd-order accurate in time, cf. Sects. 8.13.5.1 and 8.13.5.2, respectively. No upwinding is used for the explicit FE and AB predictor schemes, i.e., $\alpha = 0$. However, we note from (8.311) with $\theta = 0$ a negative numerical dispersion coefficient $D^e_{\text{num}} = -\Delta t_n q^{e2}/2$ arises for the FE predictor.

It is important to note that the stability bounds and errors of numerical dispersion listed in Table 8.9 only occur in the presence of advection $\boldsymbol{q} \neq \boldsymbol{0}$. Without advection $\boldsymbol{q} \equiv \boldsymbol{0}$ there is no need to use the PGLS, SU, SC and FU schemes.

For diffusion/conduction problems the standard GFEM is most optimal and unconditionally stable, except for the diffusion limit $Cr < Pg$, ($\Delta t_n < \acute{\mathcal{R}} h^{e2}/(2D^e)$), arising for the 2nd-order accurate TR time integration method, which is, however, commonly noncrucial.

8.16 Implementation of Dirichlet-Type BC's in the Resulting Matrix System

In formulating the weak statements in Sect. 8.9 the specification of Dirichlet boundaries Γ_D remains of particular concern. So far, the resulting weak statements do not incorporate any BC's of Dirichlet type, now we have to make up for it. In principle, there are two ways for implementing Dirichlet-type (essential) BC's. First, they can be mimicked via a Cauchy-type BC if the transfer coefficient Φ^e (or $\Phi^{\dagger^e}$) appearing in (8.104) and (8.105), associated with a global node i and the adjacent elements, is set to an arbitrary large value (theoretically, $\Phi^e \to \infty$). It enforces that the condition $\phi_i \approx \phi_C$ is satisfied in a reasonable approximation (the larger Φ^e, the better the approximation of $\phi_i \approx \phi_C$), such that the Dirichlet BC appears as a special case of the Cauchy BC. While this method is very easy and efficient, it has a clear disadvantage; namely, the large value required for the transfer coefficient increases significantly the parameter contrast in the resulting matrix system, which can deteriorate the properties of the system matrix causing negative impact on the solution of the sparse equation system.

The second method which is preferred here avoids those circumstances and satisfies the Dirichlet-type BC's in an exact manner without deteriorating the matrix system. The basic idea is that the solution at a Dirichlet boundary Γ_D is known and accordingly all nodes sharing Γ_D can be eliminated from the computations by a simple bookkeeping procedure. Let us consider a matrix system resulting from a spatio-temporal discretization such as given in (8.154) or (8.177), which leads always to a linear (or linearized) sparse system of equations written in a compact matrix form (note that the time plane indexing is dropped for convenience)

$$\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{r} \tag{8.316}$$

or

$$\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1i} & \dots & A_{1N_\mathrm{P}} \\ A_{21} & A_{22} & \dots & A_{2i} & \dots & A_{2N_\mathrm{P}} \\ \vdots & \vdots & & \vdots & & \vdots \\ A_{i1} & A_{i2} & \dots & A_{ii} & \dots & A_{iN_\mathrm{P}} \\ \vdots & \vdots & & \vdots & & \vdots \\ A_{N_\mathrm{P}1} & A_{N_\mathrm{P}2} & \dots & A_{N_\mathrm{P}i} & \dots & A_{N_\mathrm{P}N_\mathrm{P}} \end{pmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_i \\ \vdots \\ \phi_{N_\mathrm{P}} \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_i \\ \vdots \\ r_{N_\mathrm{P}} \end{pmatrix} \tag{8.317}$$

which has to be solved at each time stage for the unknown solution vector ϕ consisting of N_P components, where $\boldsymbol{A}$ is the sparse system matrix of dimension $N_\mathrm{P} \times N_\mathrm{P}$ comprising all terms of the LHS and $\boldsymbol{r}$ is the $N_\mathrm{P}-$dimensional RHS vector comprising all RHS terms of the basic discrete system (8.154) or (8.177). Now, assuming that the solution at the ith node is known, i.e., $\phi_i = \phi_D$, where ϕ_D is a prescribed Dirichlet value, then (8.317) can be rewritten

$$\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1i} & \dots & A_{1N_\mathrm{P}} \\ A_{21} & A_{22} & \dots & A_{2i} & \dots & A_{2N_\mathrm{P}} \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ A_{N_\mathrm{P}1} & A_{N_\mathrm{P}2} & \dots & A_{N_\mathrm{P}i} & \dots & A_{N_\mathrm{P}N_\mathrm{P}} \end{pmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_i \\ \vdots \\ \phi_{N_\mathrm{P}} \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ \phi_D \\ \vdots \\ r_{N_\mathrm{P}} \end{pmatrix} \tag{8.318}$$

To retrieve a possible symmetry of $\boldsymbol{A}$, it is useful to shift the ith column to the RHS. It yields the equivalent formulation

$$\begin{pmatrix} A_{11} & A_{12} & \dots 0 \dots & A_{1N_\mathrm{P}} \\ A_{21} & A_{22} & \dots 0 \dots & A_{2N_\mathrm{P}} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots 1 \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ A_{N_\mathrm{P}1} & A_{N_\mathrm{P}2} & \dots 0 \dots & A_{N_\mathrm{P}N_\mathrm{P}} \end{pmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_i \\ \vdots \\ \phi_{N_\mathrm{P}} \end{pmatrix} = \begin{pmatrix} r_1 - A_{1i}\phi_D \\ r_2 - A_{2i}\phi_D \\ \vdots \\ \phi_D \\ \vdots \\ r_{N_\mathrm{P}} - A_{N_\mathrm{P}i}\phi_D \end{pmatrix} \tag{8.319}$$

This bookkeeping procedure can be done for all Dirichlet nodes. Assuming there are N_D Dirichlet BC's in total, which are implemented in the matrix the system, the actual number of equations N_EQ which has to be solved is

$$N_\mathrm{EQ} = N_\mathrm{P} - N_\mathrm{D} \tag{8.320}$$

and the final matrix system is compressed to the actual set of equations in the form

$$\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N_\mathrm{EQ}} \\ A_{21} & A_{22} & \dots & A_{2N_\mathrm{EQ}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N_\mathrm{EQ}1} & A_{N_\mathrm{EQ}2} & \dots & A_{N_\mathrm{EQ}N_\mathrm{EQ}} \end{pmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N_\mathrm{EQ}} \end{pmatrix} = \begin{pmatrix} r_1 - A_{1i}\phi_D \\ r_2 - A_{2i}\phi_D \\ \vdots \\ r_{N_\mathrm{EQ}} - A_{N_\mathrm{EQ}i}\phi_D \end{pmatrix} \tag{8.321}$$

where all rows and columns of $\boldsymbol{A}$ are removed to which Dirichlet-type BC's are associated. In the practical solution, a *profiling* of the matrix system can be easily performed, where all Dirichlet equations are determined and eliminated from the matrix system which is actually solved. The procedure is accurate and efficient because the properties of $\boldsymbol{A}$ remain unchanged and the equation system is reduced by the N_D entries.

8.17 Solution of Linear Systems of Algebraic Equations

The spatio-temporal finite element approximation such as given by (8.154) or (8.177) leads to a matrix system in form of (8.316). After elimination the N_{D} Dirichlet-type BC's from the N_{P} equations as described in the preceding Sect. 8.16, we end up with a system of simultaneous linear (or linearized) algebraic equations written in matrix form (for sake of simplicity we drop the time plane indexing in case of transient ADE) as

$$\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b} \tag{8.322}$$

or in index notation

$$A_{ij}\phi_j = b_i \qquad (1 \leq i, j \leq N_{\mathrm{EQ}}) \tag{8.323}$$

or

$$\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N_{\mathrm{EQ}}} \\ A_{21} & A_{22} & \dots & A_{2N_{\mathrm{EQ}}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N_{\mathrm{EQ}}1} & A_{N_{\mathrm{EQ}}2} & \dots & A_{N_{\mathrm{EQ}}N_{\mathrm{EQ}}} \end{pmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N_{\mathrm{EQ}}} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{N_{\mathrm{EQ}}} \end{pmatrix} \tag{8.324}$$

which has to be solved for the $N_{\mathrm{EQ}}-$dimensional solution vector $\boldsymbol{\phi}$, where the system matrix $\boldsymbol{A}$ has N_{EQ} rows and columns. The RHS-vector $\boldsymbol{b}$, containing additionally the Dirichlet BC terms according to (8.321), has N_{EQ} components. It is to be noted that for transient problems we always prefer implicit or semi-implicit time integration schemes due to stability and performance reasons, which essentially require the solution of equation systems (in contrast to temporally explicit schemes, where in combination with mass lumping, cf. Sect. 8.13.2, there is no need to solve a system of simultaneous equations, however, at the expense of a commonly huge number of time steps as discussed in Sect. 8.13.6). Furthermore, steady-state problems, in which the time step is deemed infinitely large $\Delta t_n \rightarrow \infty$, an equation system in form of (8.322) has inevitably to be solved.

The unique solution of (8.322) at given $\boldsymbol{A}$ and $\boldsymbol{b}$ in a form

$$\boldsymbol{\phi} = \boldsymbol{A}^{-1} \cdot \boldsymbol{b} \tag{8.325}$$

only exists when $\boldsymbol{A}$ is non-singular, i.e., $\boldsymbol{A}$ must have a non-vanishing determinant $|\boldsymbol{A}| \neq 0$. The system matrix $\boldsymbol{A}$ is usually unsymmetric, i.e., $\boldsymbol{A} \neq \boldsymbol{A}^T$, when advection terms occur in the discrete formulation (except for the PGLS method introduced in Sect. 8.14.5). On the other hand, $\boldsymbol{A}$ can also be symmetric, i.e., $\boldsymbol{A} = \boldsymbol{A}^T$, for instance when terms of advection are absent. As a consequence of the used finite element discretization the system matrix $\boldsymbol{A}$ is *sparse*, i.e., many of its components are zero, and possesses a definite structure which is determined by its

Table 8.10 Advantages versus disadvantages of direct and iterative solution techniques

Method	Advantage	Disadvantage
Direct	Solution of $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$ is exact. Sequence of operations only performed once. No initial estimates and iterations are required	May be inefficient for large problems, in particular in 3D. Can produce round-off errors
Iterative	Efficient with respect to storage demand and CPU time	Solution of $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$ is approximative. Initial estimates and iteration parameters are required. System matrix $\boldsymbol{A}$ should be well-conditioned

non-zero components. A method of inversion of $\boldsymbol{A}$, in particular when the order N_{EQ} of the matrix becomes large, depends very much on the structure of $\boldsymbol{A}$. Accordingly, efficient solution methods will utilize the sparsity structure of $\boldsymbol{A}$ under exploitation of its symmetry if occurring. In general, we can differ into two major solution strategies: (1) direct and (2) iterative techniques, e.g., [15, 376, 430, 453, 590]. For large problems, particularly in 3D applications, iterative solution methods are more efficient than direct solution techniques, however, they may suffer sometimes from a poor convergence behavior. The relative merits of direct and iterative solution techniques are listed in Table 8.10. In recent years, due to the increases in computer memory and the suitability for shared-memory multiprocessing there is a revival of direct solution methods, e.g., [461].

8.17.1 Direct Solution Methods

8.17.1.1 Gaussian Elimination

The classic direct solution method is the *Gaussian elimination*. Its objective is to subtract appropriately scaled rows in the system (8.324) to arrive at an upper triangular matrix equation in the form

$$\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b} \longrightarrow \boldsymbol{U} \cdot \boldsymbol{\phi} = \boldsymbol{b}' \tag{8.326}$$

where $\boldsymbol{U}$ is an upper triangular matrix. We obtain $\boldsymbol{U}$ by following procedure termed as forward elimination: We choose the first row as the *pivot* equation and eliminate ϕ_1 from each equation below it. This is achieved by multiplying the first equation by A_{21}/A_{11} provided the *pivot element* $A_{11} \neq 0$, which is then subtracted from the second equation. It is continued similarly until ϕ_1 is eliminated from all equations. Now, we eliminate ϕ_2, ϕ_3, ... in the same manner until the upper triangular form is attained

$$\begin{pmatrix} U_{11} & U_{12} & \dots & U_{1N_{\mathrm{EQ}}} \\ 0 & U_{22} & \dots & U_{2N_{\mathrm{EQ}}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & U_{N_{\mathrm{EQ}}N_{\mathrm{EQ}}} \end{pmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N_{\mathrm{EQ}}} \end{pmatrix} = \begin{pmatrix} b'_1 \\ b'_2 \\ \vdots \\ b'_{N_{\mathrm{EQ}}} \end{pmatrix} \tag{8.327}$$

where the components of the first row are $U_{1j} = A_{1j}$, $(j{=}1,\dots,N_{\mathrm{EQ}})$ and $b'_1 = b_1$. The solution ϕ is then easily be performed by a recursive bottom-up *backsubstitution* as follows

$$\begin{aligned} \phi_{N_{\mathrm{EQ}}} &= b'_{N_{\mathrm{EQ}}} / U_{N_{\mathrm{EQ}}N_{\mathrm{EQ}}} \\ \phi_i &= (b'_i - U_{ij}\phi_j)/U_{ii}, \quad j > i \end{aligned} \tag{8.328}$$

We recognize that the Gaussian elimination changes the RHS vector $\boldsymbol{b}$ to $\boldsymbol{b}'$, which makes this technique rather inappropriate for systems with multiple RHS's. This can be circumvented by the following decomposition solution strategy.

8.17.1.2 *LU* Matrix Decomposition and Crout Method

The preferred variant of Gaussian elimination is the *Crout method*, in which the RHS vector $\boldsymbol{b}$ is not affected by the matrix decomposition. It is very advantageous for matrix systems where $\boldsymbol{A}$ does not change in time (when using constant time steps) so that $\boldsymbol{A}$ needs to be decomposed only once. In such cases ϕ can be easily computed via simple backsubstitution for every time-varying RHS vector $\boldsymbol{b}$, a considerably fast computational process termed as *resolution*. In the Crout method the matrix $\boldsymbol{A}$ is decomposed into a lower triangular matrix $\boldsymbol{L}$ and an upper triangular matrix $\boldsymbol{U}$, viz.,

$$\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{U} \tag{8.329}$$

where

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ L_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{N_{\mathrm{EQ}}1} & L_{N_{\mathrm{EQ}}2} & \dots & 1 \end{pmatrix} \tag{8.330}$$

and

$$\boldsymbol{U} = \begin{pmatrix} U_{11} & U_{12} & \dots & U_{1N_{\mathrm{EQ}}} \\ 0 & U_{22} & \dots & U_{2N_{\mathrm{EQ}}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & U_{N_{\mathrm{EQ}}N_{\mathrm{EQ}}} \end{pmatrix} \tag{8.331}$$

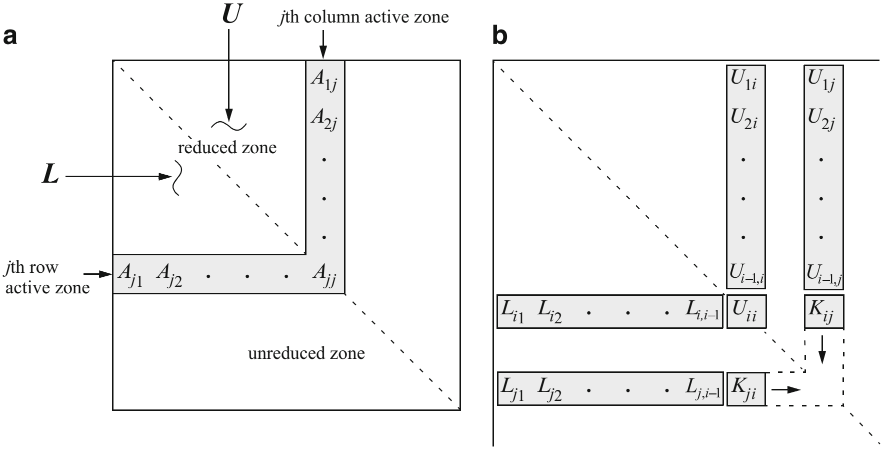


Fig. 8.35 LU decomposition of matrix $\boldsymbol{A}$ in the Crout elimination method: (**a**) reduced, active and unreduced zones, (**b**) terms used to construct U_{ij} and L_{ji}

Then, the linear equation system (8.322) expressed with (8.329)

$$\boldsymbol{A}\cdot\boldsymbol{\phi} = (\boldsymbol{L}\cdot\boldsymbol{U})\cdot\boldsymbol{\phi} = \boldsymbol{L}\cdot\underbrace{(\boldsymbol{U}\cdot\boldsymbol{\phi})}_{\boldsymbol{y}} = \boldsymbol{b} \tag{8.332}$$

can now be solved via the pair of equations

$$\boldsymbol{L}\cdot\boldsymbol{y} = \boldsymbol{b} \tag{8.333}$$

and

$$\boldsymbol{U}\cdot\boldsymbol{\phi} = \boldsymbol{y} \tag{8.334}$$

The LU decomposition (also called factorization) of $\boldsymbol{A} = \boldsymbol{L}\cdot\boldsymbol{U}$ represents the crucial and most costly solution step. The Crout method computes $\boldsymbol{L}$ and $\boldsymbol{U}$ by a continuous accumulation of products and does not need to record the intermediate reduced matrices. In this LU decomposition process the matrix $\boldsymbol{A}$ divides into three zones as outlined in Fig. 8.35. There is a region that is fully reduced, in the second (called active) zone the matrix is currently being reduced and there is a third zone, which contains the original unreduced matrix components. Taking

$$L_{ii} \equiv 1 \quad (i = 1, 2, \ldots, N_{\mathrm{EQ}}) \tag{8.335}$$

for each active zone j the entries of $\boldsymbol{U}$ and $\boldsymbol{L}$ are given by

$$\begin{aligned} U_{ij} &= A_{ij} - \sum_{m=1}^{i-1} L_{im} U_{mj} && (j = i, i+1, \ldots, N_{\mathrm{EQ}}) \\ L_{ji} &= \Big(A_{ji} - \sum_{m=1}^{i-1} L_{jm} U_{mi} \Big) / U_{ii} && (j = i+1, i+2, \ldots, N_{\mathrm{EQ}}) \end{aligned} \tag{8.336}$$

for $(i = 1, 2, \ldots, N_{\mathrm{EQ}})$. We note that the summation in (8.336) is ignored when the lower limit of the index m exceeds the upper limit.

It is obvious from (8.336) that the diagonal entry in the matrix $\boldsymbol{U}$ must be non-zero. This can be assumed for a matrix $\boldsymbol{A}$ which is diagonally dominant, i.e.,

$$|A_{ii}| \geq \sum_{j \neq i} |A_{ij}| \quad (i = 1, 2, \ldots, N_{\mathrm{EQ}}) \tag{8.337}$$

In other cases partial *pivoting* is required in which rows of $\boldsymbol{A}$ are appropriately interchanged to meet non-zero diagonals. The above decomposition can be used for both unsymmetric and symmetric matrices $\boldsymbol{A}$. However, if $\boldsymbol{A}$ is symmetric, $\boldsymbol{A} = \boldsymbol{A}^T$, the relation exists

$$U_{ij} = L_{ji} U_{ii} \tag{8.338}$$

and it is no more necessary to store the complete matrix. Only the diagonals and the components above the diagonals need to be stored, while (8.338) is utilized to construct the missing part. It reduces the decomposition costs by nearly 50 % [509].

Having completed the decomposition of $\boldsymbol{A}$, it is now trivial to solve $\boldsymbol{\phi}$ by utilizing (8.333) and (8.334) in a forward elimination and backward substitution procedure, viz.,

$$y_i = b_i - \sum_{j=1}^{i-1} L_{ij} y_j \quad (i = 1, 2, \ldots, N_{\mathrm{EQ}}) \tag{8.339}$$

and

$$\phi_i = \Big(y_i - \sum_{j=i+1}^{N_{\mathrm{EQ}}} U_{ij} \phi_j \Big) / U_{ii} \quad (i = N_{\mathrm{EQ}}, N_{\mathrm{EQ}} - 1, \ldots, 1) \tag{8.340}$$

respectively, which are computationally cheap. We see that in (8.339) the RHS vector $\boldsymbol{b}$ is not destroyed during the forward elimination, which makes the Crout method very useful for multiple RHS's in a resolution process avoiding a repeated LU decomposition.

8.17.1.3 Other Methods

A symmetric matrix $\boldsymbol{A}$ which is *positive definite*, i.e.,

$$\boldsymbol{\phi} \cdot (\boldsymbol{A} \cdot \boldsymbol{\phi}) > 0 \quad \text{for all } \boldsymbol{\phi} \neq \boldsymbol{0} \tag{8.341}$$

can be decomposed in the form

$$\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{U} = \boldsymbol{L} \cdot (\boldsymbol{D} \cdot \boldsymbol{L}^T) = \underbrace{(\boldsymbol{L} \cdot \boldsymbol{D}^{1/2})}_{\tilde{\boldsymbol{L}}} \cdot \underbrace{(\boldsymbol{D}^{1/2} \cdot \boldsymbol{L}^T)}_{\tilde{\boldsymbol{L}}^T} \tag{8.342}$$

where $\boldsymbol{D}$ is the diagonal matrix defined as

$$\boldsymbol{D} = \lceil U_{11}, U_{22}, \ldots, U_{N_{\text{EQ}}N_{\text{EQ}}} \rfloor \tag{8.343}$$

The decomposition (8.342) in the form

$$\boldsymbol{A} = \tilde{\boldsymbol{L}} \cdot \tilde{\boldsymbol{L}}^T \tag{8.344}$$

is used in the *Cholesky method*, e.g., [456,468], in which the lower triangular matrix $\tilde{\boldsymbol{L}}$ is related to $\boldsymbol{L}$ appearing in the Gaussian method by

$$\tilde{\boldsymbol{L}} = \boldsymbol{L} \cdot \boldsymbol{D}^{1/2} \tag{8.345}$$

The Cholesky method introduces a little extra-effort in computing the square root $\boldsymbol{D}^{1/2}$ compared to the Crout method for symmetric matrices. However, just this square root operation accounts for small round-off errors, which is a striking feature of the Cholesky method.

Further variants of the Gaussian elimination method differ in strategies of sparse matrix storage and bookkeeping, elimination sequences, pivoting techniques and round-off error minimizations. Active column profile solvers [509,590] based on the Crout method reduce the required storage and computational effort for unsymmetric and symmetric sparse matrices, where their columns and rows are stored only within the non-zero *profile* (also termed as envelope or skyline) of $\boldsymbol{A}$, see Fig. 8.36. It has a definite advantage over the method of a fixed banded storage. The profile can be very variable so that long and small columns can be compactly stored. The column heights are, however, dependent on the node (equation) numbering used in forming $\boldsymbol{A}$. An interesting alternative Gaussian elimination is the *frontal method* [256, 287, 291], which operates in a wave-front advancing through a finite element mesh. In contrast to a profile solution strategy, the operation sequences of the frontal method are determined by the element numbering, rather than by the node numbering. The advantage is that at no time the complete sparse matrix $\boldsymbol{A}$ must exist. Only parts of the matrix are assembled as they enter the front. However, it implies a considerable amount of bookkeeping compared to an active column profile

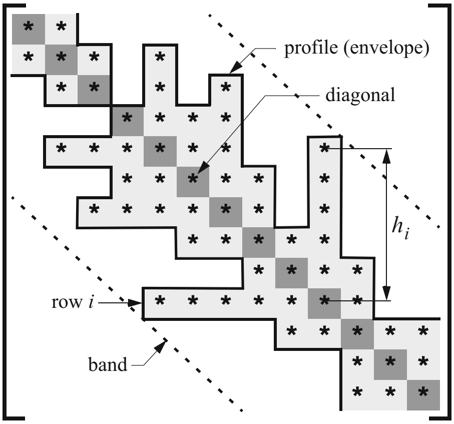


Fig. 8.36 The profile (envelope) and band of a matrix $\boldsymbol{A}$. The profile height h_i at a matrix row i is the number of columns (respectively rows) included between the first non-zero column entry and the diagonal. The bandwidth Bwd($\boldsymbol{A}$), (8.347), is formed by the maximum profile height occurring in $\boldsymbol{A}$

solver where the processing overhead remains relatively small. The frontal method is attractive in treating large matrices out-of-core, where the storage of a complete matrix would exceed the capacity of computer memory. Today, however, this is often no more a serious constraint.

8.17.1.4 Fill-in Reduction and Nodal Reordering

A profile of a sparse matrix structure as exemplified in Fig. 8.36 is formed by the non-zero entries being in a largest distance from the diagonal. This is described by the profile heights h_i $(i = 1, \ldots, N_{\text{EQ}})$ for each row i defined as

$$h_i = i - \min(j \,|\, A_{ij} \neq 0,\ j \leq i) + 1, \quad (i = 1, 2, \ldots, N_{\text{EQ}}) \tag{8.346}$$

Note that the *bandwidth* Bwd($\boldsymbol{A}$) of $\boldsymbol{A}$ is the maximum of all profile heights occurring in the matrix:

$$\text{Bwd}(\boldsymbol{A}) = \max_{1 \leq i \leq N_{\text{EQ}}} h_i \tag{8.347}$$

However, it is the nature of the finite element discretization that not all matrix entries between the first non-zero column entry and the diagonal are non-zero. Quite contrarily, a large number of entries in between can be zeros. Now, the consequence of the elimination process according to formulae (8.336) is that those zero entries within the profile of $\boldsymbol{A}$ become replaced by non-zero entries. Such entries are called *fill-in*. Since fill-in entries cause further fill-in, the complete matrix profile must be stored to perform the matrix elimination. The required storage amounts to the *envelope* (or total profile) given by the sum of the profile heights, viz.,

$$\mathrm{Env}(\boldsymbol{A}) = \begin{cases} \sum_{i=1}^{N_{\mathrm{EQ}}} h_i & \text{for symmetric } \boldsymbol{A} \\ \sum_{i=1}^{N_{\mathrm{EQ}}} (2h_i - 1) & \text{for unsymmetric } \boldsymbol{A} \end{cases} \tag{8.348}$$

On the other hand, the solution effort $\mathrm{Ecp}(\boldsymbol{A})$ of matrix elimination is proportional to the square of each profile heights

$$\mathrm{Ecp}(\boldsymbol{A}) \sim \sum_{i=1}^{N_{\mathrm{EQ}}} h_i^2 \tag{8.349}$$

In order to minimize both the storage size $\mathrm{Env}(\boldsymbol{A})$ and the computational effort $\mathrm{Ecp}(\boldsymbol{A})$ for a matrix $\boldsymbol{A}$ it is obvious that the profile heights h_i should be hold small as possible. Indeed, the profile heights are determined by the global nodal numbering used in a finite element mesh. In practical terms: the larger the difference between the highest and lowest node number occurring in a finite element, the larger the profile heights at the corresponding matrix index. This is evidenced in the example mesh shown in Fig. 8.37. While an inappropriate nodal numbering used in the mesh of Fig. 8.37a leads to a wide-spread pattern of non-zero entries in the matrix with a consequent large storage demand and a significant amount of fill-in, an intelligent nodal reordering as outlined for the mesh of Fig. 8.37b accomplishes a significant reduction of the storage demand, fill-in and computational effort.

There are different techniques [189, 418] which are useful to automatically renumber the mesh nodes with the aim to bring all matrix entries closer to the diagonal. Most important are:

- The Reverse Cuthill-McKee (RCM) method [108], which reorders the nodes according to the lowest connectivity with surrounding nodes at each level of the corresponding graph of spatial discretization.
- The Multilevel Nested Dissection (MLNDS) method [301, 302], in which the reduction of nodal interconnectivities is employed via a recursive partitioning of domains.

The RCM usually gives excellent reductions. The MLNDS is to be preferred for bigger meshes. It accomplishes a reasonable profile reduction (albeit often not so much as via RCM), however, at lower computational costs. Furthermore, MLNDS is better suitable for parallel processing. In practice, nodal reordering schemes are obligatory when direct equation solvers become in use. The nodal reordering is performed before Dirichlet-type BC's are implemented according to (8.321). It ends up with a compressed matrix system which is optimal for direct profile solvers.

8.17.2 Iterative Solution Methods

The solution of the matrix equation (8.322)

$$\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b} \tag{8.350}$$

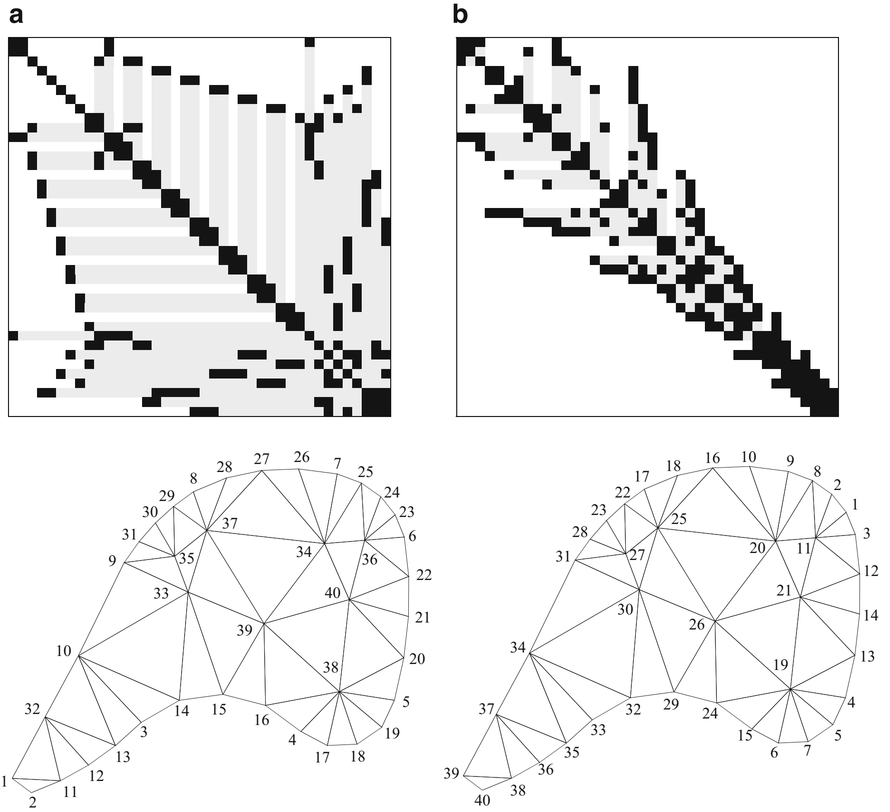


Fig. 8.37 $\boldsymbol{A}$—matrix occupations for a simple 2D triangle mesh (**a**) before and (**b**) after RCM nodal reordering. Matrix entries drawn in *black* relate to intrinsic non-zero coefficients, entries drawn in *gray* identify fill-in. The nodal reordering reduces the total profile Env($\boldsymbol{A}$) for the present mesh to about one third

by using direct methods can be rather inefficient for large systems. Their computational effort Ecp($\boldsymbol{A}$) is proportional to the square sum of all matrix profile heights (8.349), for a band structure of $\boldsymbol{A}$ it is proportional to $N_{\text{EQ}}(\text{Bwd}(\boldsymbol{A}))^2$ and for a full matrix it is even proportional to N_{EQ}^3. However, there are reliable alternatives in form of iterative solution methods, which solve (8.350) on an efficient approximate basis possessing a computational effort having only a more or less linear proportion to the equation number N_{EQ} and, however, a dependence on an iterative cycle. The faster the convergence of the iterative procedure, the smaller the required number of iterations and the better and efficient will be the iterative solution.

The principle of all iterative solution procedures is to make a first guess ϕ^0, then apply a recurrence scheme to generate a sequence of new approximations $\phi^1, \phi^2, \ldots$, that converge to ϕ. A simple recurrence scheme, known as *Richardson iteration*, could have the form $\phi^{\tau+1} = \phi^{\tau} - \chi(\boldsymbol{A} \cdot \phi^{\tau} - \boldsymbol{b})$, $(\tau = 0, 1, \ldots)$, where

τ is an iteration counter and $\chi \neq 0$ is an acceleration parameter. The advantages of such an approach are obvious: (1) the system matrix $\boldsymbol{A}$ must not be inverted directly anymore and (2) the sparsity of $\boldsymbol{A}$ can be fully exploited, where only the non-zero entries are stored in a dense manner (need not to consider fill-in). The disadvantage of iterative methods is that (1) the rate of convergence may be slow or even divergence may occur and (2) an error criteria has to be chosen at which the iteration is terminated to consider the approximate solution as sufficiently accurate. It becomes clear that the crucial point of each iteration method is to find an acceleration strategy for a fast rate of convergence. Today, there is a wide variety of iterative methods for solving both symmetric and unsymmetric systems, see e.g., [15, 453]. Most important for the present class of problems are the following:

- The Conjugate Gradient (CG) method.
- The Orthogonal Minimum Residual (ORTHOMIN) method.
- The Generalized Minimal Residual (GMRES) method.
- The Lanczos Conjugate Gradient Square (CGS) method.
- The Lanczos Bi-conjugate Gradient Stabilized (BiCGSTAB) method.
- The Multigrid (MG), in particular Algebraic Multigrid (AMG) method.

To improve the convergence behavior of these iterative methods, they are usually applied in combination with so-called preconditioning techniques which transform the basic matrix system into a form that is more suitable for the iterative procedure.

8.17.2.1 Preconditioning

An important property of the matrix $\boldsymbol{A}$ is given by the *condition number* $\kappa(\boldsymbol{A})$ defined as [453]

$$\kappa(\boldsymbol{A}) = \|\boldsymbol{A}\| \|\boldsymbol{A}^{-1}\| \tag{8.351}$$

which characterizes the ratio between the maximum and minimum eigenvalues $\kappa(\boldsymbol{A}) = \lambda_{\max}(\boldsymbol{A})/\lambda_{\min}(\boldsymbol{A})$. Problems for which κ is large are called *ill-conditioned problems*, otherwise if κ is not too large they are called *well − conditioned problems*. Typically, a high parameter contrast in the coefficients of $\boldsymbol{A}$ causes a high condition number κ. Unfortunately, the eigenvalue distribution significantly influences the convergence behavior of an iterative method. For instance in case of the CG method, if $S(P)$ is the number of iterative steps required to decrease the error $\|\phi^{\tau} - \phi\|$ by a factor of P, then $S(P) \leq \frac{1}{2}\sqrt{\kappa}\ln(2/P) + 1$. That means, the number of iterations needed to reach convergence is $\mathcal{O}(\sqrt{\kappa})$. It suggests that the rate of convergence can be significantly improved if we could decrease $\kappa \rightarrow 1$. Indeed, this is possible by a suited transformation of the basic matrix system in such a way that an iterative method will converge much faster than without this modification. Such a type of transformation is termed *preconditioning*.

To construct appropriate preconditioners for matrix $\boldsymbol{A}$, we differ between explicit and implicit preconditioning methods. To solve $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$, an explicit method transforms the system into

$$(\boldsymbol{C}^{-1} \cdot \boldsymbol{A}) \cdot \boldsymbol{\phi} = \boldsymbol{C}^{-1} \cdot \boldsymbol{b} \tag{8.352}$$

where $\boldsymbol{C}$ is the preconditioning matrix to be chosen. Then, (8.352) is solved on an iterative basis, e.g.,

$$\boldsymbol{\phi}^{\tau+1} = (\boldsymbol{\delta} - \tilde{\boldsymbol{A}}) \cdot \boldsymbol{\phi}^{\tau} + \tilde{\boldsymbol{b}}, \quad \text{where} \quad \tilde{\boldsymbol{A}} = \boldsymbol{C}^{-1} \cdot \boldsymbol{A} \quad \text{and} \quad \tilde{\boldsymbol{b}} = \boldsymbol{C}^{-1} \cdot \boldsymbol{b} \tag{8.353}$$

In implicit preconditioning the original problem $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$ is replaced by a sequence of solutions of the form

$$\boldsymbol{C} \cdot (\boldsymbol{\phi}^{\tau+1} - \boldsymbol{\phi}^{\tau}) = -\boldsymbol{r}^{\tau} \tag{8.354}$$

with the residual vector

$$\boldsymbol{r}^{\tau} = \boldsymbol{A} \cdot \boldsymbol{\phi}^{\tau} - \boldsymbol{b} \tag{8.355}$$

Both forms (8.352) and (8.354) are equivalent. Their use is only dependent on how the inverse $\boldsymbol{C}^{-1}$ is explicitly known. In any cases, the preconditioning matrix $\boldsymbol{C}$ should require only little computational extra-effort. On the other hand, the chosen preconditioning matrix must significantly improve the eigenspectrum of $\boldsymbol{C}^{-1} \cdot \boldsymbol{A}$ in comparison to $\boldsymbol{A}$, i.e., $\kappa(\boldsymbol{C}^{-1} \cdot \boldsymbol{A}) < \kappa(\boldsymbol{A})$. Both requirements are somewhat contradictory: the better the preconditioning, the higher often the costs. Note that for the case $\boldsymbol{C} = \boldsymbol{A}$ in (8.354) the scheme corresponds to a direct solution and the sequence of solutions stops after one iteration. In approximating $\boldsymbol{A}$ with $\boldsymbol{C}$ it tends to drop low eigenvalues (i.e., long-wavelength eigenmodes), what can be a deficiency. Hence, a good and optimal choice of $\boldsymbol{C}$ (or explicitly $\boldsymbol{C}^{-1}$) is desired. Today, a large family of preconditioners is available, cf. [15, 45, 453]. Our preferred conditioning methods are:

- The incomplete LU (ILU) decomposition method: $\boldsymbol{C} = \boldsymbol{L} \cdot \boldsymbol{U}$ for unsymmetric and $\boldsymbol{C} = \boldsymbol{L} \cdot (\boldsymbol{D} \cdot \boldsymbol{L}^T)$ for symmetric systems.
- The modified ILU (MILU) method, where diagonal entries of $\boldsymbol{U}$ are additionally modified to tackle ill-conditioned problems.
- The polynomial preconditioning: $\boldsymbol{C}^{-1} = p(\boldsymbol{A})$, where $p(\boldsymbol{A})$ is a polynomial of lower degree, commonly Chebyshev polynomials.

Most useful is the incomplete lower-upper (ILU) preconditioning in the form

$$\boldsymbol{C} = \begin{cases} \boldsymbol{L} \cdot \boldsymbol{U} & \text{unsymmetric} \\ \boldsymbol{L} \cdot (\boldsymbol{D} \cdot \boldsymbol{L}^T) & \text{symmetric} \end{cases} \tag{8.356}$$

which is achieved by a matrix decomposition with a Crout method (8.336), however, the fill-in that occurs for all the off-diagonals within the matrix profile is completely or partly neglected. The simplest and usually preferred method is the ILU decomposition (factorization) with *no fill-in*, termed by ILU(0). It works very fast, is robust and needs only a comparatively small extra-storage. For certain applications an extended ILU preconditioning can be suitable in which some fill-in is allowed in the incomplete LU decomposition, e.g., ILU(1) which accomplishes 1st-order fill-ins, see [453] for more. Commonly, nodal reordering (see Sect. 8.17.1.4) is not needed for ILU(0), but in using ILU(1) it can improve the accuracy of the preconditioner due to the fill-in-minimized rearranged structure of the matrix.

Improvements of ILU preconditioning can be attained by so-called modified ILU (MILU) preconditioners [40]. Basically, these techniques are zero fill-in ILU(0) strategies, however, the fill-in entries occurring during the matrix decomposition process are kept (e.g., adding up positive off-diagonal entries) in order to put the lumped sum to the diagonal entries. In this way one attempts to compensate the discarded entries, which can be important for the lower eigenvalues if the matrix is ill-conditioned. A favorite is the Gustaffson MILU preconditioning [217], which is designed and specialized for symmetric matrices including a high coefficient contrast. However, the MILU strategy has shown often insuffiently robust and should not be applied in general.

8.17.2.2 The Preconditioned Conjugate Gradient (PCG) Method

The conjugate gradient (CG) method goes back to Hestenes and Stiefel [247], who presented a new iterative method with a significantly increased rate of convergence in solving sparse *symmetric* positive-definite equation systems. Bizarrely, it took many years until this powerful method has found acceptance in the numerical analysis community that were exclusively fixed on direct solution methods over long time. But, beginning in the 1970s and in particular once computers became powerful enough to tackle real 3D problems, the CG method has started its triumph and gained considerable attraction in numerical modeling. Today, the CG method has become the standard iterative method for sparse symmetric equation systems. Its major advantages are: (1) the number of operations per iterative step τ is only proportional to the number of equations N_{EQ} and (2) it converges in at most N_{EQ} iterations in the absence of round-off errors. In practice, however, the method already converges after a relatively small number of iterations much faster than the pessimistic estimate of N_{EQ}. The rate of convergence depends on the distribution of eigenvalues. Accordingly, the use of appropriate preconditioning further increases the rate of convergence of the CG method. The *preconditioned CG (PCG) method* is the preferred iterative solution method for symmetric matrix systems.

The iterative algorithm for solving $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$ by the PCG method is given as follows (see e.g., [15, 453]):

Let $\boldsymbol{A}$ and $\boldsymbol{C}$ be symmetric and positive definite. Guess initially $\boldsymbol{\phi}^0$
and set: $\boldsymbol{r}^0 = \boldsymbol{A}\cdot\boldsymbol{\phi}^0 - \boldsymbol{b}$, $\boldsymbol{h}^0 = \boldsymbol{C}^{-1}\cdot\boldsymbol{r}^0$ and $\boldsymbol{d}^0 = -\boldsymbol{h}^0$,
with known values: ϵ and ITMAX

For iterations $\tau = 0, 1, 2, \ldots$ compute until convergence:

$$\begin{aligned} \boldsymbol{\phi}^{\tau+1} &= \boldsymbol{\phi}^{\tau} + \alpha^{\tau}\boldsymbol{d}^{\tau} \qquad \text{where} \quad \alpha^{\tau} = \frac{\boldsymbol{h}^{\tau^T}\cdot\boldsymbol{r}^{\tau}}{\boldsymbol{d}^{\tau^T}\cdot(\boldsymbol{A}\cdot\boldsymbol{d}^{\tau})} \\ \boldsymbol{r}^{\tau+1} &= \boldsymbol{r}^{\tau} + \alpha^{\tau}(\boldsymbol{A}\cdot\boldsymbol{d}^{\tau}) \\ \boldsymbol{h}^{\tau+1} &= \boldsymbol{C}^{-1}\cdot\boldsymbol{r}^{\tau+1} \\ \boldsymbol{d}^{\tau+1} &= -\boldsymbol{h}^{\tau+1} + \beta^{\tau}\boldsymbol{d}^{\tau} \qquad \text{where} \quad \beta^{\tau} = \frac{\boldsymbol{h}^{\tau+1^T}\cdot\boldsymbol{r}^{\tau+1}}{\boldsymbol{h}^{\tau^T}\cdot\boldsymbol{r}^{\tau}} \end{aligned} \tag{8.357}$$

Stop if $\frac{\boldsymbol{r}^{\tau+1^T}\cdot\boldsymbol{r}^{\tau+1}}{\boldsymbol{b}^T\cdot\boldsymbol{b}} < \epsilon^2$ *or* $\tau > \mathit{ITMAX}$

where $\boldsymbol{r}$ is the residual vector, $\boldsymbol{h}$ is the pseudoresidual vector, $\boldsymbol{d}$ is the search direction vector, $\boldsymbol{C}$ is the preconditioning matrix, commonly $\boldsymbol{C} = \boldsymbol{L}\cdot(\boldsymbol{D}\cdot\boldsymbol{L}^T)$ by using an ILU(0) preconditioner, ϵ is the termination criterion (default 10^{-8}) and *ITMAX* is the allowed maximum number of iterations (e.g., 200) to be chosen in dependence on N_{EQ}.

8.17.2.3 The Preconditioned Restarted ORTHOMIN Method

The orthogonal minimum residual (ORTHOMIN) method belongs to a family of generalized conjugate gradient methods. It was firstly presented by Vinsome [547] and widely used in petroleum reservoir simulation. A biorthogonal vector algorithm, originally attributed to Lanczos [332], forms the basis for solving sparse unsymmetric matrix systems. Most of the following variants of iterative methods applied to unsymmetric matrices are based on that biorthogonalization procedure, termed as *Lanczos algorithm*.

In iteratively solving $\boldsymbol{A}\cdot\boldsymbol{\phi} = \boldsymbol{b}$ by the ORTHOMIN method, the orthogonality of $\boldsymbol{A}\cdot\boldsymbol{q}^{\tau}$ is required, that is $(\boldsymbol{A}\cdot\boldsymbol{q}^{\tau})\cdot(\boldsymbol{A}\cdot\boldsymbol{q}^{k}) = 0$ for $\tau \neq k$, where $\boldsymbol{q}^{\tau}$ and $\boldsymbol{q}^{k}$ are the search directions at iterative steps τ and k, respectively. The ORTHOMIN method converges to the exact solution $\boldsymbol{\phi}$ within N_{EQ} iterations, however, it is necessary to store up N_{EQ} search directions $\boldsymbol{q}^{\tau}$ and N_{EQ} products $\boldsymbol{A}\cdot\boldsymbol{q}^{\tau}$. This implies large summation which makes the procedure sensitive for accumulating round-off errors in the computation of the search directions. To overcome this deficiency the orthogonalization is restarted every K iterations, i.e., the ORTHOMIN procedure runs for K steps to get an approximation $\boldsymbol{\phi}^{K}$, then setting $\boldsymbol{\phi}^0 = \boldsymbol{\phi}^{K}$ and restart the iterations until convergence is reached. This is referred to as *restarted* ORTHOMIN or ORTHOMIN(K), e.g., [342]. Since the computational cost increases as $\mathcal{O}(K^2 N_{\mathrm{EQ}})$ and the memory cost increases as $\mathcal{O}(K N_{\mathrm{EQ}})$, it is required to hold K small relative to N_{EQ}. Usually, K is chosen in the range between 4 and 10, depending inversely on the condition number of the matrix. In practice, it has been shown that the restarted ORTHOMIN converges in a similar number of iterations as the non-restarted version [42].

The iterative algorithm for solving $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$ by the preconditioned restarted ORTHOMIN(K) method is given as follows (e.g., [346, 547]):

```
Let $\boldsymbol{A}$ and $\boldsymbol{C}$ be unsymmetric. Guess initially $\boldsymbol{\phi}^0$ and set: $\boldsymbol{r}^0 = \boldsymbol{b} - \boldsymbol{A} \cdot \boldsymbol{\phi}^0$,
$i = 0$, with the known values: $\epsilon$, ITMAX and $K$

(1) For iterations $\tau = 1, 2, \ldots, K$ do:
    $i = i + 1$
    $\boldsymbol{u}^\tau = \boldsymbol{C}^{-1} \cdot \boldsymbol{r}^{\tau-1}$
    $\boldsymbol{v}^\tau = \boldsymbol{A} \cdot \boldsymbol{u}^\tau$
    $\boldsymbol{p}^\tau = \boldsymbol{u}^\tau$
    $\boldsymbol{q}^\tau = \boldsymbol{v}^\tau$
    For $1 \le k \le \tau$, do $\left\{ \begin{array}{l} \alpha^{k\tau} = (\boldsymbol{q}^{k^T} \cdot \boldsymbol{v}^\tau)/(\boldsymbol{q}^{k^T} \cdot \boldsymbol{q}^k) \\ \boldsymbol{p}^\tau = \boldsymbol{p}^\tau - \alpha^{k\tau} \boldsymbol{p}^k \\ \boldsymbol{q}^\tau = \boldsymbol{q}^\tau - \alpha^{k\tau} \boldsymbol{q}^k \end{array} \right.$
    $\beta^\tau = (\boldsymbol{q}^{\tau^T} \cdot \boldsymbol{r}^{\tau-1})/(\boldsymbol{q}^{\tau^T} \cdot \boldsymbol{q}^\tau)$
    $\boldsymbol{\phi}^\tau = \boldsymbol{\phi}^{\tau-1} + \beta^\tau \boldsymbol{p}^\tau$
    $\boldsymbol{r}^\tau = \boldsymbol{r}^{\tau-1} - \beta^\tau \boldsymbol{q}^\tau$
    Compute $e_1^\tau = \frac{\|\boldsymbol{r}^\tau\|_{L_\infty}}{\|\boldsymbol{b}\|_{L_\infty}}$, $e_2^\tau = \frac{\beta^\tau \|\boldsymbol{p}^\tau\|_{L_\infty}}{\|\boldsymbol{\phi}^\tau\|_{L_\infty}}$
    Stop if $e_1^\tau < \epsilon$ or $e_2^\tau < \epsilon$ or $i \ge$ ITMAX
(2) End do
    $\boldsymbol{r}^0 = \boldsymbol{r}^K$
    $\boldsymbol{\phi}^0 = \boldsymbol{\phi}^K$
Go to (1) for restarting
```

(8.358)

where $\boldsymbol{r}$ is the residual vector, $\boldsymbol{q}$ is the search direction vector, $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{p}$ are auxiliary vectors, $\boldsymbol{C}$ is the preconditioning matrix, preferentially $\boldsymbol{C} = \boldsymbol{L} \cdot \boldsymbol{U}$ by using an ILU(0) Crout preconditioner, ϵ is the termination criterion (default 10^{-6}), $\|.\|_{L_\infty}$ is the maximum norm defined by (8.26), K is the number of iterations (default 5) after which the algorithm is periodically restarted and *ITMAX* is the tolerated maximum of total iterations (e.g., 200) to be chosen in dependence on N_{EQ}.

The ORTHOMIN(K) method is only guaranteed to converge for *positive real* matrices $\boldsymbol{A}$, that is if $\boldsymbol{\phi} \cdot (\boldsymbol{A} \cdot \boldsymbol{\phi}) > 0$ for all $\boldsymbol{\phi} \neq \boldsymbol{0}$. In other cases, there is no guarantee anymore for convergence, unless an appropriate preconditioning matrix $\boldsymbol{C}$ can be found which creates a positive real matrix $\boldsymbol{C}^{-1} \cdot \boldsymbol{A}$.

8.17.2.4 The Preconditioned Restarted GMRES Method

The generalized minimal residual (GMRES) method was introduced by Saad and Schultz [454] for solving unsymmetric matrix systems. It is mathematically equivalent to a generalized conjugate gradient method, however, the orthogonal vector algorithm is based on the Arnoldi method (see e.g., [15, 453]), which saves computational effort and improves robustness in particular for large equation systems. Li et al. [346] have shown that GMRES can be one-third faster than ORTHOMIN in various large-scale petroleum reservoir applications. GMRES is guaranteed to converge in at most N_{EQ} steps, provided that $\boldsymbol{A}$ or $\boldsymbol{C}^{-1}\cdot\boldsymbol{A}$ is positive real. However, similar to the ORTHOMIN method the storage demand of GMRES increases linearly with the iteration τ and the number of operations increases as $\mathcal{O}(\tau^2 N_{\text{EQ}})$, which is rather computationally impractical for large matrices. The alternative is that GMRES is to restart after a fixed number of iterations K, similar to a restarted ORTHOMIN(K) procedure. In GMRES(K), the GMRES method is periodically restarted after every K iterations until reaching convergence.

The iterative algorithm for solving $\boldsymbol{A}\cdot\boldsymbol{\phi} = \boldsymbol{b}$ by the preconditioned restarted GMRES(K) method is given as follows (e.g., [342, 346, 453, 454]):

```
Let $\boldsymbol{A}$ and $\boldsymbol{C}$ be unsymmetric. Choose a first guess $\phi^0$. Set up the
$(K+1)\times K$ Hessenberg matrix: $\boldsymbol{H}^K = (H^{k\tau})_{1\le k\le K+1, 1\le\tau\le K} = \mathbf{0}$, $i = 0$,
with the known values: $\epsilon_1$, $\epsilon_2$, ITMAX and $K$

(1) Arnoldi process:
Compute $\boldsymbol{r}^0 = \boldsymbol{C}^{-1}\cdot(\boldsymbol{b} - \boldsymbol{A}\cdot\phi^0)$, $\beta = \|\boldsymbol{r}^0\|_{L_2}$ and $\boldsymbol{v}^1 = \boldsymbol{r}^0/\beta$
For iterations $\tau = 1, 2, \ldots, K$ do:
  $i = i + 1$
  $\boldsymbol{w}^\tau = \boldsymbol{C}^{-1}\cdot(\boldsymbol{A}\cdot\boldsymbol{v}^\tau)$
  For $1\le k\le\tau$, do if $(\boldsymbol{w}^{\tau^T}\cdot\boldsymbol{v}^k) > \epsilon_1\|\boldsymbol{w}^\tau\|_{L_2}$: $\left\{\begin{array}{l} H^{k\tau} = \boldsymbol{w}^{\tau^T}\cdot\boldsymbol{v}^k \\ \boldsymbol{w}^\tau = \boldsymbol{w}^\tau - H^{k\tau}\boldsymbol{v}^k \end{array}\right.$
  $H^{\tau+1,\tau} = \|\boldsymbol{w}^\tau\|_{L_2}$
  $\boldsymbol{v}^{\tau+1} = \boldsymbol{w}^\tau/\|\boldsymbol{w}^\tau\|_{L_2}$
End do
Define $\boldsymbol{V}^K = (\boldsymbol{v}^1, \ldots, \boldsymbol{v}^K)^T$ and update $\phi^K = \phi^0 + \boldsymbol{V}^K y^K$,
where $y^K$ is the minimizer of $\|\beta\boldsymbol{e}_1 - \boldsymbol{H}^K y\|_{L_2}$ with $\boldsymbol{e}_1 = (1, 0, 0, \ldots, 0)^T$
Stop if $\frac{\|\boldsymbol{r}^K\|_{L_2}}{\|\boldsymbol{b}\|_{L_2}} \le \epsilon_2$ or $i \ge$ ITMAX, where $\boldsymbol{r}^K = \boldsymbol{C}^{-1}\cdot(\boldsymbol{b} - \boldsymbol{A}\cdot\phi^K)$,
otherwise set $\phi^0 = \phi^K$ and go to (1) for restarting
```

(8.359)

where $\boldsymbol{r}$ is the residual vector, $\boldsymbol{H}$ is the Hessenberg matrix [453], $\boldsymbol{v}$ and $\boldsymbol{w}$ are auxiliary vectors, $\boldsymbol{C}$ is the preconditioning matrix, preferentially $\boldsymbol{C} = \boldsymbol{L}\cdot\boldsymbol{U}$ by using an ILU(0) Crout preconditioner, ϵ_1 is the criterion for checking orthogonality (default 10^{-10}), ϵ_2 is the convergence criterion to terminate the iteration (default 10^{-6}), $\|.\|_{L_2}$ is the L_2 error norm defined by (8.25), K is the number of iterations (default 5) after which the algorithm is periodically restarted and *ITMAX* is the tolerated maximum of total iterations (e.g., 200) to be chosen in dependence on N_{EQ}.

8.17.2.5 The Preconditioned Lanczos Conjugate Gradient Square (CGS) Method

The CGS method is a variant of the Lanczos-type biorthogonalization (biconjugate gradient) method for solving unsymmetric matrix systems in which biorthogonal sets of vectors are generated. It has been proposed by Sonneveld [487]. CGS is a highly efficient iterative method for unsymmetric matrices which converges about twice as fast than standard biconjugate gradient methods. It is based on squaring the residual polynomials. While CGS works well in many applications, it is prone to rounding errors due to the squared polynomials and even breakdowns cannot be fully precluded in cases causing a divide by zero. On the other hand, the residual error is not strictly decreasing in the progress of iterations. Nevertheless, its properties regarding the smallest storage demand and accelerated convergence in the most applications make the CGS method in combination with an appropriate preconditioning a powerful recurrence scheme for solving unsymmetric equation systems.

The iterative algorithm for solving $\boldsymbol{A}\cdot\boldsymbol{\phi} = \boldsymbol{b}$ by the preconditioned CGS method can be written in the following form (e.g., [453, 487]):

Let $\boldsymbol{A}$ *and* $\boldsymbol{C}$ *be unsymmetric. Guess initially* $\boldsymbol{\phi}^0$
and set: $\boldsymbol{r}^0 = \boldsymbol{q}^0 = \boldsymbol{C}^{-1}\cdot(\boldsymbol{b} - \boldsymbol{A}\cdot\boldsymbol{\phi}^0),\ \boldsymbol{g}^0 = \boldsymbol{h}^0 = \boldsymbol{0},\ \beta^0 = \boldsymbol{q}^{0^T}\cdot\boldsymbol{r}^0$ *and*
$\gamma^0 = 0$ *with known values:* ϵ *and ITMAX*

For iterations $\tau = 0, 1, 2, \ldots$ *compute until convergence:*

$$\begin{aligned}
\boldsymbol{y}^{\tau+1} &= \boldsymbol{r}^{\tau} + \gamma^{\tau}\boldsymbol{h}^{\tau}\\
\boldsymbol{g}^{\tau+1} &= \boldsymbol{y}^{\tau+1} + \gamma^{\tau}(\boldsymbol{h}^{\tau} + \gamma^{\tau}\boldsymbol{g}^{\tau})\\
\boldsymbol{h}^{\tau+1} &= \boldsymbol{y}^{\tau+1} - \alpha^{\tau+1}\boldsymbol{C}^{-1}\cdot(\boldsymbol{A}\cdot\boldsymbol{g}^{\tau+1}) \quad \text{where} \quad \alpha^{\tau+1} = \frac{\beta^{\tau}}{\boldsymbol{q}^{0^T}\cdot[\boldsymbol{C}^{-1}\cdot(\boldsymbol{A}\cdot\boldsymbol{g}^{\tau+1})]}\\
\boldsymbol{\phi}^{\tau+1} &= \boldsymbol{\phi}^{\tau} + \alpha^{\tau+1}(\boldsymbol{y}^{\tau+1} + \boldsymbol{h}^{\tau+1})\\
\boldsymbol{r}^{\tau+1} &= \boldsymbol{r}^{\tau} - \alpha^{\tau+1}\boldsymbol{C}^{-1}\cdot[\boldsymbol{A}\cdot(\boldsymbol{y}^{\tau+1} + \boldsymbol{h}^{\tau+1})]\\
\beta^{\tau+1} &= \boldsymbol{q}^{0^T}\cdot\boldsymbol{r}^{\tau+1}\\
\gamma^{\tau+1} &= \beta^{\tau+1}/\beta\tau
\end{aligned}$$

Stop if $\frac{\|\boldsymbol{r}^{\tau+1}\|_{L_2}}{\|\boldsymbol{r}^0\|_{L_2}} < \epsilon$ *or* $\tau >$ *ITMAX*

(8.360)

where $\boldsymbol{r}$ is the residual vector, $\boldsymbol{q}$ is the shadow residual vector, $\boldsymbol{y}$, $\boldsymbol{h}$, $\boldsymbol{g}$ are auxiliary vectors, $\boldsymbol{C}$ is the preconditioning matrix, preferentially $\boldsymbol{C} = \boldsymbol{L}\cdot\boldsymbol{U}$ by using an ILU(0) Crout preconditioner, ϵ is the termination criterion (default 10^{-8}) and *ITMAX* is the allowed maximum number of iterations (e.g., 200) to be chosen in dependence on N_{EQ}. The preconditioned CGS method converges in at most N_{EQ} iterations for positive real matrices.

8.17.2.6 The Preconditioned Lanczos Bi-conjugate Gradient Stabilized (BiCGSTAB) Method

Van der Vorst [535] proposed the BiCGSTAB algorithm as an improved variant of CGS. The BiCGSTAB method stabilizes and smoothes the convergence behavior. It leads to a more robust and usually faster converging iterative technique for solving unsymmetric equations systems with small storage demand and low computational cost. The BiCGSTAB iteration steps are only slightly more expensive than the CGS steps. But, similar to CGS the BiCGSTAB method cannot fully exclude the risk of a computational breakdown if the system matrix $\boldsymbol{A}$ is not positive real. In most applications, however, BiCGSTAB has shown a superior behavior by what it has become a preferred iterative method for solving unsymmetric matrix systems.

The iterative algorithm for solving $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$ by the preconditioned BiCGSTAB method is given as follows (e.g., [453, 535]):

Let $\boldsymbol{A}$ *and* $\boldsymbol{C}$ *be unsymmetric. Guess initially* $\boldsymbol{\phi}^0$
and set: $\boldsymbol{r}^0 = \boldsymbol{b} - \boldsymbol{A} \cdot \boldsymbol{\phi}^0$ *and* $\boldsymbol{p}^0 = \boldsymbol{q}^0 = \boldsymbol{C}^{-1} \cdot \boldsymbol{r}^0$
with known values: ϵ *and ITMAX*

For iterations $\tau = 0, 1, 2, \ldots$ *compute until convergence:*

$$\begin{aligned}
\boldsymbol{s}^\tau &= \boldsymbol{r}^\tau - \alpha^\tau \boldsymbol{A} \cdot \boldsymbol{p}^\tau \qquad \text{where} \quad \alpha^\tau = \frac{\boldsymbol{q}^{0^T} \cdot (\boldsymbol{C}^{-1} \cdot \boldsymbol{r}^\tau)}{\boldsymbol{q}^{0^T} \cdot [\boldsymbol{C}^{-1} \cdot (\boldsymbol{A} \cdot \boldsymbol{p}^\tau)]} \\
\boldsymbol{t}^\tau &= \boldsymbol{C}^{-1} \cdot \boldsymbol{s}^\tau \\
\boldsymbol{\phi}^{\tau+1} &= \boldsymbol{\phi}^\tau + \alpha^\tau \boldsymbol{p}^\tau + \omega^\tau \boldsymbol{t}^\tau \quad \text{where} \quad \omega^\tau = \frac{(\boldsymbol{A} \cdot \boldsymbol{t}^\tau)^T \cdot \boldsymbol{s}^\tau}{(\boldsymbol{A} \cdot \boldsymbol{t}^\tau)^T \cdot (\boldsymbol{A} \cdot \boldsymbol{t}^\tau)} \\
\boldsymbol{r}^{\tau+1} &= \boldsymbol{s}^\tau - \omega^\tau \boldsymbol{A} \cdot \boldsymbol{t}^\tau \\
\beta^\tau &= \frac{\alpha^\tau}{\omega^\tau} \, \frac{\boldsymbol{q}^0 \cdot (\boldsymbol{C}^{-1} \cdot \boldsymbol{r}^{\tau+1})}{\boldsymbol{q}^0 \cdot (\boldsymbol{C}^{-1} \cdot \boldsymbol{r}^\tau)} \\
\boldsymbol{p}^{\tau+1} &= \boldsymbol{C}^{-1} \cdot \boldsymbol{r}^{\tau+1} + \beta^\tau [\boldsymbol{p}^\tau - \omega^\tau \boldsymbol{C}^{-1} \cdot (\boldsymbol{A} \cdot \boldsymbol{p}^\tau)] \\
&\textit{Stop if} \quad \frac{\|\boldsymbol{r}^{\tau+1}\|_{L_2}}{\|\boldsymbol{r}^0\|_{L_2}} < \epsilon \quad \textit{or} \quad \tau > \textit{ITMAX}
\end{aligned} \tag{8.361}$$

where $\boldsymbol{r}$ is the residual vector, $\boldsymbol{q}$ is the shadow residual vector, $\boldsymbol{s}$, $\boldsymbol{t}$, $\boldsymbol{p}$ are auxiliary vectors, $\boldsymbol{C}$ is the preconditioning matrix, preferentially $\boldsymbol{C} = \boldsymbol{L} \cdot \boldsymbol{U}$ by using an ILU(0) Crout preconditioner, ϵ is the termination criterion (default 10^{-8}) and *ITMAX* is the allowed maximum number of iterations (e.g., 200) to be chosen in dependence on N_{EQ}. The preconditioned BiCGSTAB method converges in at most N_{EQ} iterations for positive real matrices.

8.17.2.7 Multigrid (MG) Methods

The iterative solution methods treated so far suffer from disabling limitations: (1) Their convergence rates are dependent on the number of equations N_{EQ} to be solved. This has severe implications in the numerical solution of very large problems involving millions and billions of mesh nodes. The required number of iterations inevitably increases and can reach an unacceptable size. (2) Their convergence rate

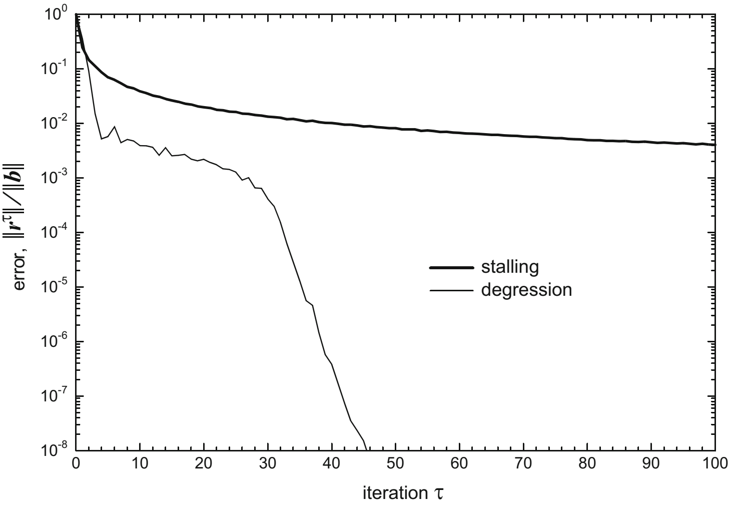


Fig. 8.38 Convergence behavior of PCG method showing a typical well-behaved degression of residual error and the occurrence of stalling in a large problem solution

has a tendency to stall, in particular for large problems, that means the reduction rate of errors becomes slow or even practically stagnant. In fact, these iterative methods converge very rapidly for the few iterations and very slowly thereafter, see Fig. 8.38. The reason for that unfavorable behavior is obvious: The convergence rate is a function of the error field frequency, i.e., the measure of change of the error from node to node. All high error frequencies or small wavelength components which are comparable to the mesh size can be effectively reduced (smoothed out), see Fig. 8.39, however, low error frequencies or large wavelength components of error can only badly annihilated such that the convergence rate automatically deteriorates. As the mesh is refined, the low error frequencies dominate the solution error and additional iterations become progressively less productive. Indeed, this represents a serious limitation of those iterative methods. But, the remedy is possible by using multigrid (MG) methods.

The basic idea of MG methods is likewise simple and intuitive: Since low frequency errors remain widely hidden for fine grids (meshes), it should be more efficient to reduce those errors on coarser grids (Fig. 8.40). In using both fine and coarse grids in an appropriate interplay it must per se lead to a highly powerful iterative strategy, where both high and low error frequencies are reduced at the same time with fast convergence. The natural way of transfering between fine and coarser grids firstly results in the traditional MG method, called geometric multigrid (GMG).

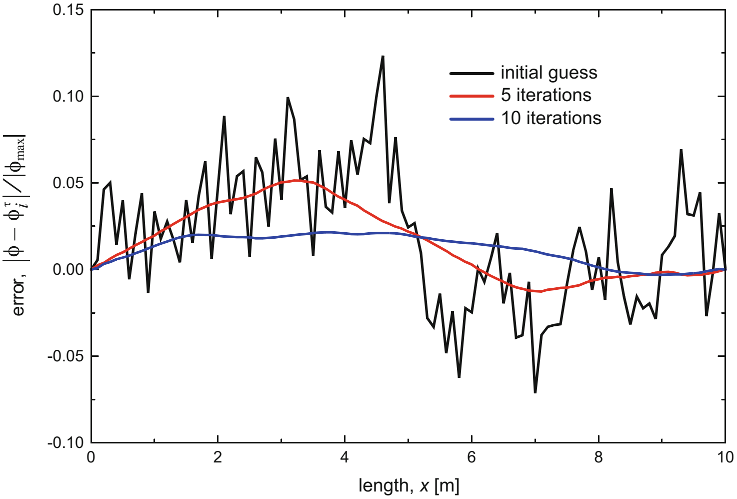


Fig. 8.39 Error $|\phi - \phi_i^\tau|/|\phi_{\max}|$ $(i = 1, \ldots, N_P)$ reduction (smoothing) of PCG method in the course of iterations τ applied to a 1D diffusion problem

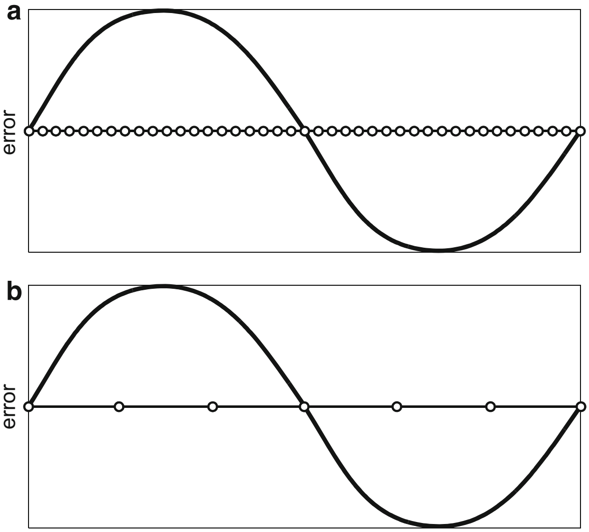


Fig. 8.40 (**a**) On a fine mesh, the error is visible as low frequency and large wavelength, (**b**) on a coarse mesh, the error is seen as high(er) frequency and small(er) wavelength. Smooth modes on a fine grid look less smooth on a coarse grid

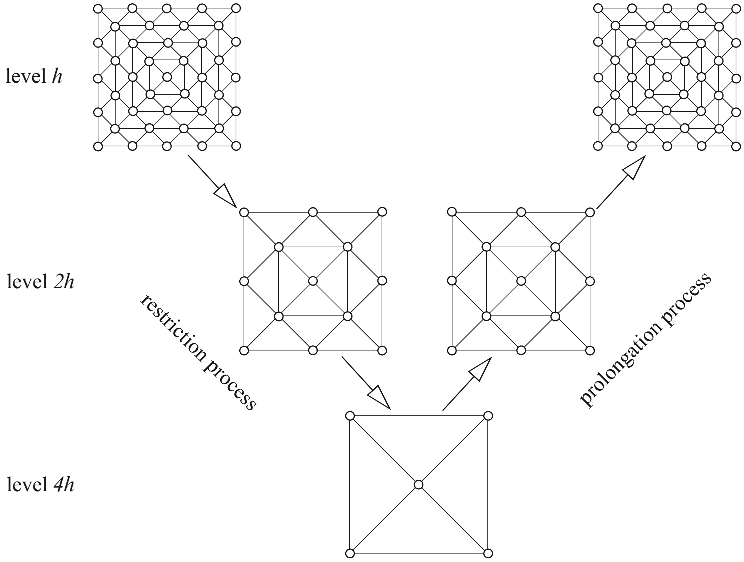


Fig. 8.41 Example of a three-level multigrid hierarchy consisting of uniform triangle meshes

Geometric Multigrid (GMG) Method

The concept of the multigrid algorithm dates back to the 1960s when Fedorenko [161] and Bakhvalov [21] published their first studies. Other multigrid pioneers are Brandt [51] and Hackbusch [219], who have recognized in the 1970s the actual efficiency of the MG method and started fundamental developments. Today, for MG methods an extensive mathematical basis exists and a variety of efficient numerical strategies for many applications have been worked out. A good overview is given in the textbook by Trottenberg et al. [519]. The classic and standard MG approach refers to the geometric MG (GMG) method.

The usual practice in GMG consists of a successive *nested* structured grid (mesh) procedure in which the coarse grid has twice the grid spacing $2h$ of the next finer grid with the grid size h so that all nodes in the coarse grid also appear in the fine grid. The use of grid spacings with a ratio of 2 allows very efficient intergrid transfer operations and lead to *hierarchical* meshes where typically fine mesh elements result from coarse mesh elements by a simple subdivision via element halving [219]. A hierarchical mesh can also be locally refined. An example for a three-level successive nested multigrid hierarchy consisting of uniform triangle meshes is exhibited in Fig. 8.41 which features a V-cycle.

To illustrate the GMG procedure for solving the finite element matrix system $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$, the sequence of only two mesh levels identified by subscripts h for the fine grid and $2h$ for the next coarse grid are considered at first. It begins with solving $\boldsymbol{A}_h \cdot \boldsymbol{\phi}_h = \boldsymbol{b}_h$ on the fine mesh for a small number of iterations <10 by using an appropriate iterative method (e.g., PCG or others) until the residual

$$\boldsymbol{r}_h^\tau = \boldsymbol{b}_h - \boldsymbol{A}_h \cdot \boldsymbol{\phi}_h^\tau \tag{8.362}$$

is sufficiently smooth, i.e., high error frequencies are suitably reduced. This solution step is termed *presmoothing*. Then, to effectively reduce low error frequencies the so-called *coarse grid correction* starts. It transfers the residual $\boldsymbol{r}_h^\tau$ to the coarse grid, a process called *restriction*, in the following form of a coarse grid defect matrix equation

$$\boldsymbol{A}_{2h} \cdot \Delta\boldsymbol{\phi}_{2h}^\tau = \boldsymbol{b}_{2h}^\tau \quad \text{with} \quad \boldsymbol{b}_{2h}^\tau = \boldsymbol{I}_h^{2h} \cdot \boldsymbol{r}_h^\tau \quad \text{and} \quad \Delta\boldsymbol{\phi}_{2h}^\tau = \boldsymbol{\phi}_{2h}^{\tau+1} - \boldsymbol{\phi}_{2h}^\tau \tag{8.363}$$

where $\boldsymbol{I}_h^{2h}$ is a nonsquare matrix, known as the *restriction operator*. The solution increment $\Delta\boldsymbol{\phi}_{2h}^\tau$ results from (8.363) and can now be transferred back to the fine grid by interpolation, a process called *prolongation*, to obtain

$$\Delta\boldsymbol{\phi}_h^\tau = \boldsymbol{I}_{2h}^h \cdot \Delta\boldsymbol{\phi}_{2h}^\tau \tag{8.364}$$

where $\boldsymbol{I}_{2h}^h$ is a nonsquare matrix, known as the *prolongation operator*. It gives the new approximation

$$\boldsymbol{\phi}_h^{\tau+1} = \boldsymbol{\phi}_h^\tau + \Delta\boldsymbol{\phi}_h^\tau \tag{8.365}$$

on the fine mesh. A *postsmoothing* solution step can now follow in which the residual $\boldsymbol{r}_h^{\tau+1} = \boldsymbol{b}_h - \boldsymbol{A}_h \cdot \boldsymbol{\phi}_h^{\tau+1}$ is further reduced via a standard iterative solver to obtain an improved solution $\boldsymbol{\phi}_h^{\tau+1}$ on the fine mesh.

The construction of the restriction operator $\boldsymbol{I}_h^{2h}$ and the prolongation operator $\boldsymbol{I}_{2h}^h$ can be rather simple when using nested grids in which all coarse grid nodes appear in the fine grid nodes. In the FEM context the natural choice is the use of interpolations based on the basis functions (8.16) such that

$$\sum_j N_{2h,j}\, \phi_{2h,j} \approx \sum_l N_{h,l}\, \phi_{h,l} \tag{8.366}$$

where $N_{2h,j}$ and $N_{2h,l}$ denote the basis functions of meshes $2h$ with nodes j and h with nodes l, respectively. To minimize the approximation error a Galerkin-weighting approach for (8.366) becomes useful

$$\begin{aligned} \textstyle\int_\Omega N_{2h,i} N_{2h,j}\, d\Omega\, \phi_{2h,j} &= \textstyle\int_\Omega N_{2h,i} N_{h,l}\, d\Omega\, \phi_{h,l} \\ \boldsymbol{O}_{2h} \cdot \boldsymbol{\phi}_{2h} &= \boldsymbol{M}_h^{2h} \cdot \boldsymbol{\phi}_h \end{aligned} \tag{8.367}$$

where $\boldsymbol{O}_{2h}$ is a consistent mass matrix for the coarse mesh, which can also be lumped (see Sect. 8.13.2), and $\boldsymbol{M}_h^{2h}$ forms a new integral that consists of the inner product of basis functions from the different meshes. The restriction operator directly follows from (8.367) as

$$\boldsymbol{I}_h^{2h} = \boldsymbol{O}_{2h}^{-1} \cdot \boldsymbol{M}_h^{2h} \tag{8.368}$$

which is simply evaluable for the lumped matrix $\boldsymbol{O}_{2h}^{-1} = \boldsymbol{\delta}$. A similar relation can be developed for the prolongation operator. We recognize that (8.368) implies operations in form of averages between adjacent grid points. More difficulties arise in using unnested and unstructured grids, see e.g., [354], where the operators have to be carefully derived to avoid additional inacceptable approximation errors appearing in the intergrid interpolations.

The principle of the two-grid procedure as stated above can be generalized to sequences of multiple grids. The reason for using more sequences of grids is obvious: The solution of (8.363) on the coarse grid may not be much different from the next fine grid. Hence, we can recursively repeat this two-grid procedure on successively coarser grids, creating coarser and coarser grids, down to some coarsest grid. Then, on the coarsest grid the remaining defect equation of the type (8.363) can be usually solved exactly via a direct solver. The solution is then prolongated successively to the finer grids. The multigrid algorithm takes the form:

$$\begin{array}{l}
\text{smooth } \boldsymbol{r}_h^\tau = \boldsymbol{b}_h - \boldsymbol{A}_h \cdot \boldsymbol{\phi}_h^\tau \quad n \text{ times} \\
\text{restrict } \boldsymbol{b}_{2h}^\tau = \boldsymbol{I}_h^{2h} \cdot \boldsymbol{r}_h^\tau \\
\qquad \text{smooth } \boldsymbol{r}_{2h}^\tau = \boldsymbol{b}_{2h}^\tau - \boldsymbol{A}_{2h} \cdot \Delta\boldsymbol{\phi}_{2h}^\tau \quad n \text{ times} \\
\qquad \text{restrict } \boldsymbol{b}_{4h}^\tau = \boldsymbol{I}_{2h}^{4h} \cdot \boldsymbol{r}_{2h}^\tau \\
\qquad\qquad \text{smooth } \boldsymbol{r}_{4h}^\tau = \boldsymbol{b}_{4h}^\tau - \boldsymbol{A}_{4h} \cdot \Delta\boldsymbol{\phi}_{4h}^\tau \quad n \text{ times} \\
\qquad\qquad \text{restrict } \boldsymbol{b}_{8h}^\tau = \boldsymbol{I}_{4h}^{8h} \cdot \boldsymbol{r}_{4h}^\tau \\
\qquad\qquad\qquad \vdots \\
\qquad\qquad \text{prolongate } \Delta\boldsymbol{\phi}_{4h}^\tau = \boldsymbol{I}_{8h}^{4h} \cdot \Delta\boldsymbol{\phi}_{8h}^\tau \\
\qquad\qquad \text{smooth } \boldsymbol{r}_{4h}^\tau = \boldsymbol{b}_{4h}^\tau - \boldsymbol{A}_{4h} \cdot \Delta\boldsymbol{\phi}_{4h}^\tau \quad m \text{ times} \\
\qquad \text{prolongate } \Delta\boldsymbol{\phi}_{2h}^\tau = \boldsymbol{I}_{4h}^{2h} \cdot \Delta\boldsymbol{\phi}_{4h}^\tau \\
\qquad \text{smooth } \boldsymbol{r}_{2h}^\tau = \boldsymbol{b}_{2h}^\tau - \boldsymbol{A}_{2h} \cdot \Delta\boldsymbol{\phi}_{2h}^\tau \quad m \text{ times} \\
\text{prolongate } \Delta\boldsymbol{\phi}_h^\tau = \boldsymbol{I}_{2h}^{h} \cdot \Delta\boldsymbol{\phi}_{2h}^\tau \\
\text{compute } \boldsymbol{\phi}_h^{\tau+1} = \boldsymbol{\phi}_h^\tau + \Delta\boldsymbol{\phi}_h^\tau \\
\text{smooth } \boldsymbol{r}_h^{\tau+1} = \boldsymbol{b}_h - \boldsymbol{A}_h \cdot \boldsymbol{\phi}_h^{\tau+1} \quad m \text{ times to finalize} \quad \boldsymbol{\phi}_h^{\tau+1}
\end{array} \tag{8.369}$$

encompassing *one* iteration step (cycle) τ of the multigrid procedure. Such a consecutive fine-to-coarse and coarse-to-fine multigrid cycle is called V-cycle, sketched in Fig. 8.42. However, there are much more options of how to cycle the multiple grids. Another possibility is the W-cycle, where more coarse grids are visited to drive the residuals down as much as possible before returning to the more expensive finer grids (Fig. 8.42). In cases where the initial solution on the fine mesh may be too poor, a full multigrid cycle (Fig. 8.42) is appropriate to obtain better starting solutions on the coarse grids.

It can be shown for the GMG method [519] that its *convergence is independent of the size of the finest grid.* Solving a problem in D dimensions, the reduction in the number of nodal points N_P between subsequent grids is of the order of $N_\mathrm{P}^{2h}/N_\mathrm{P}^{h} \sim 1/2^D$. Assuming that n smoothing steps are required on each grid and the computational work of each smoothing processes is proportional to the effective

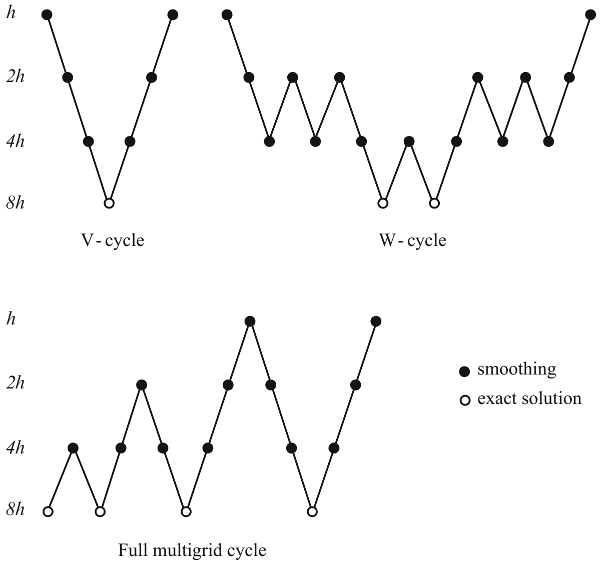


Fig. 8.42 Types of multigrid cycle exemplified for a four-grid method

number of equations N_{EQ}, the total computational work needed for a full V-cycle is only of the order $\mathcal{O}(n\ N_{\mathrm{EQ}}^{h} \log_D(N_{\mathrm{EQ}}^{h}))$ and the associated storage requirement only of the order $\mathcal{O}(N_{\mathrm{EQ}}^{h} \log_D(N_{\mathrm{EQ}}^{h}))$. In fact, these estimates are extremely favorable and hardly to be beaten by any other iterative strategy. For a W-cycle the amount of work is only slightly larger. In particular for 3D problems, we observe that the number of grid points on the coarser grids drops dramatically.

Algebraic Multigrid (AMG) Method

While the GMG method has shown very efficient in particular for large and very large problems, there are unfortunately a number of serious deficiencies which hamper its use in the finite element modeling practice. Detrimental is that GMG often deteriorates for problems with anisotropic and discontinuous coefficients. More important is that GMG depends fundamentally on the availability of an underlying grid. The treatment of complex meshes in 3D has shown often rather cumbersome. The FEM generally uses unstructured, nonhierarchical meshes. For that mesh complexity it is difficult if not impossible to construct reliable GMG methods. However, the basic principles of GMG can be exploited in a generalized strategy without suffering from GMG's fundamental restrictions. Such a strategy has become true with the *algebraic multigrid* (AMG) method [453, 519].

Brandt [52] and Stüben [497] can be seen as the major protagonists of the AMG method, who started AMG's development in the early 1980s. It was motivated by the observation that straightforward geometric grid transfer operations of restriction and prologation can be alternatively formulated on the basis of the underlying matrices without any reference to grids (meshes), i.e., the construction of these operators can be done purely algebraically. On the other hand, AMG's algorithmic components of smoothing and coarse-grid correction remain completely analogous to the classical GMG method maintaining its computational power and efficiency for solving large matrix systems. In contrast to GMG where coarse-grid discretizations are used to reduce low-frequency error components, the AMG method reduces the low error frequencies on matrix equations of a reduced dimension defining a certain level. As a consequence, AMG does not require anymore fixed grid hierarchies. Accordingly, in AMG one should better use the term *multilevel* rather than multigrid (but for historical reasons the term multigrid is often further preferred in the AMG context).

AMG is best developed for scalar elliptic PDE, however, recent progress has also been attained for systems of PDE's and ADE's. It has been proven to be a very robust and efficient solution method applicable to both structured and unstructured meshes. Stüben [498] gives a comprehensive overview on AMG in different fields of application. The key feature of AMG is the exploitation of the Galerkin approach comparable to (8.367) of GMG, however, in the context of AMG the required coarse-grid operators of restriction and prolongation are based on interpolation that maps a coarse node into a fine one of a given matrix system. This coarsening process is fully automatic. Suppose the finite element matrix system $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$ is given at the highest level (finest mesh)

$$\boldsymbol{A}_h \cdot \boldsymbol{\phi}_h = \boldsymbol{b}_h \tag{8.370}$$

similar to a geometric two-grid description, we can define the matrix system for the next coarse-level problem identified by subscript H as

$$\boldsymbol{A}_H \cdot \boldsymbol{\phi}_H = \boldsymbol{b}_H \tag{8.371}$$

The coarse-level AMG system (8.371) is constructed by means of the Galerkin approach. In doing so, the coarse matrix $\boldsymbol{A}_H$ results from

$$\boldsymbol{A}_H = (\boldsymbol{I}_h^H \cdot \boldsymbol{A}_h) \cdot \boldsymbol{I}_H^h \tag{8.372}$$

where $\boldsymbol{I}_h^H$ and $\boldsymbol{I}_H^h$ denote the restriction and prolongation operators, respectively. Their construction forms the major task of AMG's setup process, see [498] for more details. Having these operators a two-level process and by its recursive application any multilevel process can be easily performed in an analogy to GMG's multigrid cycle and smoothing algorithms described above.

8.18 Treatment of Nonlinearities

In the previous Sect. 8.17 we have described techniques for solving the resulting algebraic equations systems $\boldsymbol{A} \cdot \boldsymbol{\phi} = \boldsymbol{b}$, provided that the system is *linear*, i.e., the solution $\boldsymbol{\phi}$ does not occur in any other combination than in a linear one. However, in a number of applications the parameters in the governing finite element equations can contain dependencies on the solution $\boldsymbol{\phi}$ itself. Typical examples are variable density and variable saturation problems, where the advective and diffusive (conductive) terms in $\boldsymbol{A}$ become a function of $\boldsymbol{\phi}$. Furthermore, nonlinear BC's and higher order reaction processes imply nonlinear dependencies in the RHS vector $\boldsymbol{b}$. In such cases a nonlinear matrix system results in a form

$$\boldsymbol{A}(\boldsymbol{\phi}) \cdot \boldsymbol{\phi} = \boldsymbol{b}(\boldsymbol{\phi}) \tag{8.373}$$

where the main nonlinear functional dependence is identified by parentheses. Before we can solve (8.373) for $\boldsymbol{\phi}$ the system of equations must be *linearized* by using appropriate iterative methods. Most important are the Picard iteration method, which is a linearly convergent algorithm, and the Newton iteration method, which normally converges quadratically. A specific concern is suitable for transient problems.

8.18.1 Fixed Point Form and Picard Iteration Method

The system of nonlinear equations (8.373) written as

$$\boldsymbol{R}(\boldsymbol{\phi}) = \boldsymbol{0} \quad \text{with} \quad \boldsymbol{R}(\boldsymbol{\phi}) = \boldsymbol{A}(\boldsymbol{\phi}) \cdot \boldsymbol{\phi} - \boldsymbol{b}(\boldsymbol{\phi}) \tag{8.374}$$

can be given in its *fixed point form* as

$$\boldsymbol{\phi} = \boldsymbol{G}(\boldsymbol{\phi}) \quad \text{with} \quad \boldsymbol{G}(\boldsymbol{\phi}) = \boldsymbol{A}^{-1}(\boldsymbol{\phi}) \cdot \boldsymbol{b}(\boldsymbol{\phi}) \tag{8.375}$$

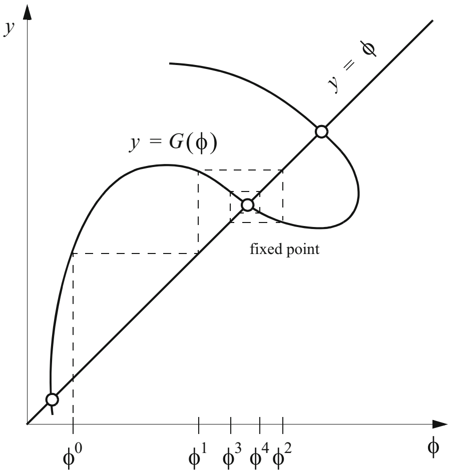
Solutions of (8.375) are called fixed points of the mapping function $\boldsymbol{G}(\boldsymbol{\phi})$, which represent solutions of (8.374). In graphical terms, fixed points are the intersections of the graph $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\phi})$ with the line $\boldsymbol{y} = \boldsymbol{\phi}$ as illustrated in Fig. 8.43 for a scalar functional dependence.

The fixed point form (8.375) immediately suggests the following iteration scheme

$$\boldsymbol{\phi}^{\tau+1} = \boldsymbol{G}(\boldsymbol{\phi}^{\tau}) \quad \tau = 0, 1, 2, \ldots \tag{8.376}$$

where τ is the iteration counter. The formulation (8.376) is the method of *successive substitution* known as the *Picard iteration method*. The iteration is started with a first

Fig. 8.43 Fixed points of $y = G(\phi)$ and fixed point (Picard) iteration starting from the initial value ϕ^0



guess ϕ^0 so that $G(\phi^0)$ can be evaluated to obtain ϕ^1. Repeating this procedure a sequence of successive solutions for $\phi^{\tau+1}$ is obtained (see illustration in Fig. 8.43). In the practical application, however, the matrix G is not directly formed. Instead, the Picard iteration is executed in the basic matrix system in the form

$$\boldsymbol{A}(\boldsymbol{\phi}^{\tau}) \cdot \boldsymbol{\phi}^{\tau+1} = \boldsymbol{b}(\boldsymbol{\phi}^{\tau}) \quad \tau = 0, 1, 2, \ldots \tag{8.377}$$

to obtain $\phi^{\tau+1}$. The iteration method *linearizes* the matrix system so that the equation system (8.377) can be easily solved by using solution techniques of Sect. 8.17. An advantage of the Picard method is that the structural matrix properties remain unchanged, in particular, if $\boldsymbol{A}$ in (8.373) is *symmetric* the matrix system (8.377) remains symmetric. During the iterative loop the matrix system $\boldsymbol{A}$ and the RHS $\boldsymbol{b}$ must be updated (reassembled) with the previous solution and the equation system (8.377) has to be repeatedly solved until satisfactory convergence is achieved. A typical convergence criterion is

$$\frac{\|\boldsymbol{\phi}^{\tau+1} - \boldsymbol{\phi}^{\tau}\|}{\|\boldsymbol{\phi}^{\tau+1}\|} \leq \epsilon \tag{8.378}$$

where ϵ is an error tolerance to be prescribed and $\|.\|$ corresponds to a suitable error norm, e.g., RMS error norm (8.28) or maximum error norm (8.29).

The proof of the convergence for the Picard iteration is given by the *Banach fixed point theorem*, e.g., [199]. It is shown that the iteration error of the Picard method decreases *linearly* with the error of the previous iteration step, viz.,

$$\|\boldsymbol{\phi}^{\tau+1} - \boldsymbol{\phi}^{\tau}\| < \|\boldsymbol{\phi}^{\tau} - \boldsymbol{\phi}^{\tau-1}\| \quad \text{for} \quad \tau > 0 \tag{8.379}$$

provided that the initial estimate for the solution ϕ^0 is within a contracting distance

$$\|\phi^0 - \phi\| \leq r_c \tag{8.380}$$

to converge to the unique solution ϕ, where r_c is referred to as the *radius of convergence* of the Picard iteration scheme. In general, for the present class of problems it is not possible to determine r_c. For strong nonlinearities the convergence radius can be very small and a good first guess of the solution is usually needed to attain a converging solution, otherwise the method diverges and fails. Nevertheless, the Picard iteration method has shown relatively robust in many applications. Its robustness is however paid by an only linear (1st-order) convergence rate.

8.18.2 Newton Iteration Method

The *Newton iteration method*, also known as the *Newton-Raphson method*, possesses a more rapid convergence behavior in form of a quadratic convergence rate. Considering the nonlinear matrix equation (8.374) written as

$$\boldsymbol{R}(\boldsymbol{\phi}) = \boldsymbol{A}(\boldsymbol{\phi}) \cdot \boldsymbol{\phi} - \boldsymbol{b}(\boldsymbol{\phi}) = \boldsymbol{0} \tag{8.381}$$

and assuming that the residual $\boldsymbol{R}(\boldsymbol{\phi})$ is continuous and differentiable, a Taylor series expansion for the residual at the new iteration $\boldsymbol{R}(\boldsymbol{\phi}^{\tau+1})$ about the previous iterative solution $\boldsymbol{\phi}^{\tau}$ yields

$$\boldsymbol{R}(\boldsymbol{\phi}^{\tau+1}) = \boldsymbol{R}(\boldsymbol{\phi}^{\tau}) + \frac{\partial \boldsymbol{R}(\boldsymbol{\phi}^{\tau})}{\partial \boldsymbol{\phi}^{\tau}} \cdot \Delta\boldsymbol{\phi}^{\tau} + \mathcal{O}(\Delta\boldsymbol{\phi}^{\tau^2}) + \ldots \tag{8.382}$$

with

$$\Delta\boldsymbol{\phi}^{\tau} = \boldsymbol{\phi}^{\tau+1} - \boldsymbol{\phi}^{\tau} \tag{8.383}$$

Assuming $\boldsymbol{R}(\boldsymbol{\phi}^{\tau+1}) = \boldsymbol{0}$ and neglecting 2nd and higher order terms, we obtain from (8.382) the Newton iteration scheme in the form

$$\boldsymbol{J}(\boldsymbol{\phi}^{\tau}) \cdot \Delta\boldsymbol{\phi}^{\tau} = -\boldsymbol{R}(\boldsymbol{\phi}^{\tau}) \quad \tau = 0, 1, 2, \ldots \tag{8.384}$$

where

$$\boldsymbol{R}(\boldsymbol{\phi}^{\tau}) = \boldsymbol{A}(\boldsymbol{\phi}^{\tau}) \cdot \boldsymbol{\phi}^{\tau} - \boldsymbol{b}(\boldsymbol{\phi}^{\tau}) \tag{8.385}$$

and the tangential matrix

$$\boldsymbol{J}(\boldsymbol{\phi}^{\tau}) = \frac{\partial \boldsymbol{R}(\boldsymbol{\phi}^{\tau})}{\partial \boldsymbol{\phi}^{\tau}} = \frac{\partial}{\partial \boldsymbol{\phi}^{\tau}}[\boldsymbol{A}(\boldsymbol{\phi}^{\tau}) \cdot \boldsymbol{\phi}^{\tau} - \boldsymbol{b}(\boldsymbol{\phi}^{\tau})] = \boldsymbol{A}(\boldsymbol{\phi}^{\tau}) + \hat{\boldsymbol{J}}(\boldsymbol{\phi}^{\tau}) \tag{8.386}$$

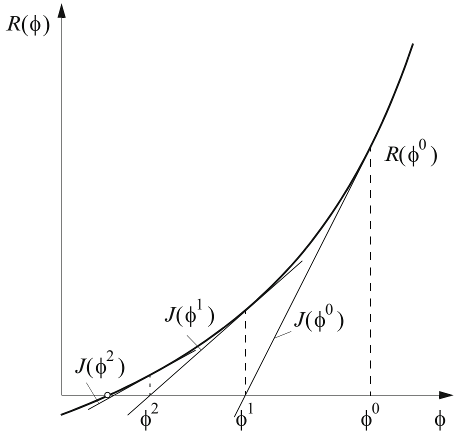


Fig. 8.44 Illustration of Newton iteration method for solving the nonlinear function $R(\phi) = 0$. Case of convergence. Iteration starts from the initial value ϕ^0. The Jacobians $J(\phi^\tau) = \partial R(\phi^\tau)/\partial\phi^\tau$ represent the tangential slopes which are variable for each iteration $\tau = 0, 1, 2, \ldots$

with

$$\hat{\boldsymbol{J}}(\phi^\tau) = \frac{\partial \boldsymbol{A}(\phi^\tau)}{\partial \phi^\tau} \cdot \phi^\tau - \frac{\partial \boldsymbol{b}(\phi^\tau)}{\partial \phi^\tau} \tag{8.387}$$

in which $\boldsymbol{J}(\phi^\tau)$ and $\hat{\boldsymbol{J}}(\phi^\tau)$ are the Jacobian matrix and the *partial* Jacobian matrix, respectively. It is important to note that the Jacobian has to be updated for each iteration (*full Newton method*) to realize the quadratic convergence rate as illustrated in Fig. 8.44 for a scalar functional dependence. On the other hand, the partial Jacobian $\hat{\boldsymbol{J}}$ causes always an *unsymmetric* matrix even if the system matrix $\boldsymbol{A}$ is symmetric. For these reasons the Newton iteration method is relatively expensive. The partial Jacobian $\hat{\boldsymbol{J}}$ can be computed either analytically or numerically.[16] Usually, the analytical evaluation is preferred because it provides a more efficient implementation. From (8.386) we can recognize that the full Newton method reduces to the Picard method (8.377) when the partial Jacobian $\hat{\boldsymbol{J}}$ (8.387) is dropped such that $\boldsymbol{J}(\phi^\tau) \approx \boldsymbol{A}(\phi^\tau)$.

For terminating the Newton iteration scheme (8.384) the deviatory convergence criterion of (8.378) may be applied. However, the convergence of the Newton

[16] If the Jacobian $\boldsymbol{J}(\phi^\tau) = \partial\boldsymbol{R}(\phi^\tau)/\partial\phi^\tau$ is not analytically available or too difficult for an analytical evaluation, it can be constructed numerically via a secant approximation by using a possibly very small increment δ in a form such as

$$\boldsymbol{J}(\phi^\tau) \approx \frac{\boldsymbol{R}(\phi^\tau + \delta) - \boldsymbol{R}(\phi^\tau)}{\delta}$$

The increment δ should not be chosen too small to avoid roundoff errors. On the other hand, a too large δ leads to a poor approximation of the Jacobian. A reasonable choice is the square root of the unit roundoff being about $\epsilon_R = 10^{-12}$ in double precision arithmetic, accordingly $\delta = \sqrt{\epsilon_R} = 10^{-6}$. The extra effort of the numerical evaluation consists of additional N_{EQ} evaluation of residual $\boldsymbol{R}$.

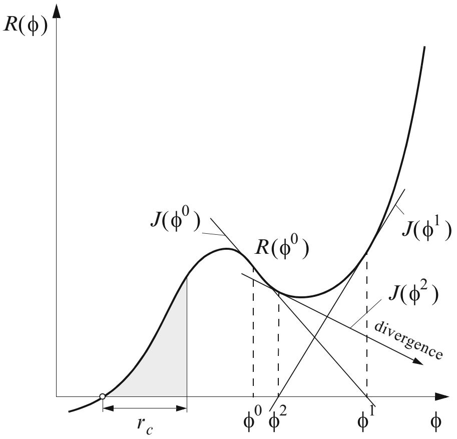


Fig. 8.45 Diverging Newton iteration in solving the nonlinear function $R(\phi) = 0$ if initial value ϕ^0 is outside the convergence radius r_c

method can easily (and additionally) be controlled by another useful error criterion, viz., the test of the minimal residual, e.g., written in the form

$$\frac{\|\boldsymbol{R}(\phi^{\tau+1})\|}{\|\boldsymbol{F}(\phi^{\tau+1})\|} \leq \epsilon_2 \tag{8.388}$$

normalized for instance by the RHS vector $\boldsymbol{F}$ (appearing in (8.154) or (8.177)), where ϵ_2 represents a second convergence criterion. The advantage of this test is that the global balance error of the spatio-temporal matrix system is directly controlled. The acceptable measure of the minimal residual ϵ_2 can be chosen suitably small, possibly in the range of the roundoff error.

It is known that the Newton method requires a good first guess of the solution ϕ^0, otherwise if the starting solution is too far from the correct solution the method can 'blow up' and quickly diverges (see Fig. 8.45). The convergence radius r_c as defined in (8.380) is generally smaller for the Newton method than for the Picard iteration method. It is complicated further in the Newton method that r_c decreases as the number of equations N_{EQ} increases so that ϕ^0 must be closer to the correct solution for bigger meshes. There are various cost-effective modifications in the Newton iteration method to reduce the increased computational effort in updating the Jacobian, where a concomitantly slower convergence rate (commonly linear) has to be accepted. Most important are the modified Newton method and the quasi-Newton method introduced next.

8.18.3 Modified Newton and Quasi-Newton Iteration Method

A major drawback of the full Newton method is that the Jacobian $\boldsymbol{J}(\phi^\tau)$ has to be updated in each iteration τ. Although its quadratic convergence leads usually

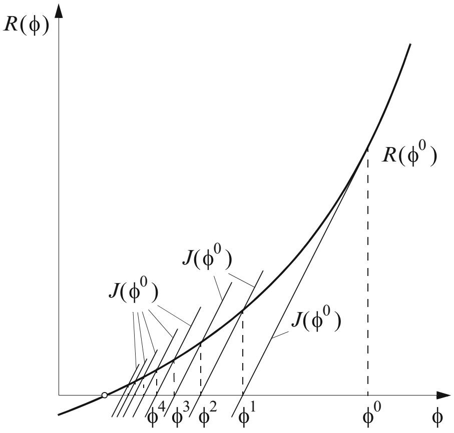


Fig. 8.46 Illustration of modified Newton iteration method for solving the nonlinear function $R(\phi) = 0$. Case of convergence. Iteration starts from the initial value ϕ^0. The Jacobian $J(\phi^0) = \partial R(\phi^0)/\partial\phi^0$ represents the initial tangential slope at ϕ^0 which is kept and used in all subsequent iterations

to a small number of iterations, each iteration is accordingly expensive. There are variants of the Newton method which can result in fewer costly iterations, however, at the expense of a slower convergence. Nevertheless, since each iterative step is ostensibly cheaper one can afford more iterations.

To obviate the need of Jacobian updating the *modified Newton method* can be used in which the Jacobian is only built once at the initial step, i.e., $\boldsymbol{J}(\phi^0)$ is formed with the initial solution ϕ^0. Then, all subsequent iterations leave this initial Jacobian $\boldsymbol{J}(\phi^0)$ unchanged (see Fig. 8.46), i.e.,

$$\boldsymbol{J}(\phi^0) \cdot \Delta\phi^\tau = -\boldsymbol{R}(\phi^\tau) \quad \text{with} \quad \boldsymbol{J}(\phi^0) = \frac{\partial \boldsymbol{R}(\phi^0)}{\partial \phi^0} \tag{8.389}$$

The convergence rate of the modified Newton iteration method is only linear. However, in comparison to the Picard method, which is also linearly convergent, the modified Newton algorithm is usually cheaper because it needs only one matrix update per iteration cycle.

Another possible cost-effective modification is the *quasi-Newton method*. In this case the Jacobian $\boldsymbol{J}(\phi^\tau)$ can be thought of as approximations to the system matrix $\boldsymbol{A}(\phi^\tau)$. The quasi-Newton iteration can be written in the form

$$\begin{aligned} \phi^{\tau+1} &= \phi^\tau - s^\tau \boldsymbol{A}^{-1}(\phi^\tau) \cdot \boldsymbol{R}(\phi^\tau) \\ \boldsymbol{A}^{-1}(\phi^{\tau+1}) &= \boldsymbol{A}^{-1}(\phi^\tau) + \Delta\boldsymbol{A}^{-1}(\phi^\tau) \end{aligned} \tag{8.390}$$

where $\boldsymbol{A}(\phi^\tau)$ has to satisfy the secant condition

$$\boldsymbol{A}(\phi^\tau) \cdot (\phi^\tau - \phi^{\tau-1}) = \boldsymbol{R}(\phi^\tau) - \boldsymbol{R}(\phi^{\tau-1}) \tag{8.391}$$

in which s^τ is an acceleration factor (usually, $s^\tau = 1$) and the update of the system matrix is expressed directly via an incremental correction $\Delta\boldsymbol{A}^{-1}(\phi^\tau)$ to its inverse.

The efficiency of the quasi-Newton method is dependent on finding good choices of inverse update forms. For a symmetric matrix $\boldsymbol{A}$ the *Broyden-Fletcher-Goldfarb-Shannon (BFGS)* update [121] has shown most successful. Broyden's update is also available for unsymmetric matrices, for more see, e.g., [156]. The quasi-Newton method possesses usually a better than linear convergence rate and is accordingly superior to the modified Newton method.

8.18.4 Transient Nonlinear Problem Solution

For transient problems the nonlinear matrix system has to be solved at the time plane $n+1$:

$$\boldsymbol{A}(\phi_{n+1}) \cdot \phi_{n+1} = \boldsymbol{b}(\phi_{n+1}) \tag{8.392}$$

In accordance with the used time integration methods different strategies have found appropriate. In principle, at each time plane the nonlinear system (8.392) must be iteratively solved to achieve convergence. The iteration procedure reads for the Picard method

$$\boldsymbol{A}(\phi_{n+1}^{\tau}) \cdot \phi_{n+1}^{\tau+1} = \boldsymbol{b}(\phi_{n+1}^{\tau}) \quad \tau = 0, 1, 2, \ldots \tag{8.393}$$

and for the full Newton method

$$\begin{aligned} \boldsymbol{J}(\phi_{n+1}^{\tau}) \cdot \Delta\phi_{n+1}^{\tau} &= -\boldsymbol{R}(\phi_{n+1}^{\tau}) \quad \tau = 0, 1, 2, \ldots \\ \Delta\phi_{n+1}^{\tau} &= \phi_{n+1}^{\tau+1} - \phi_{n+1}^{\tau} \\ \boldsymbol{J}(\phi_{n+1}^{\tau}) &= \frac{\partial \boldsymbol{R}(\phi_{n+1}^{\tau})}{\partial \phi_{n+1}^{\tau}} \\ \boldsymbol{R}(\phi_{n+1}^{\tau}) &= \boldsymbol{A}(\phi_{n+1}^{\tau}) \cdot \phi_{n+1}^{\tau} - \boldsymbol{b}(\phi_{n+1}^{\tau}) \end{aligned} \tag{8.394}$$

The iteration usually starts at time plane $n+1$ with the first guess taking from the previous time n, i.e., $\phi_{n+1}^{0} = \phi_n$. The process is repeated within each time plane until the following convergence criteria are satisfied:

$$\frac{\|\phi_{n+1}^{\tau+1} - \phi_{n+1}^{\tau}\|}{\|\phi_{n+1}^{\tau+1}\|} \leq \epsilon \tag{8.395}$$

and/or

$$\frac{\|\boldsymbol{R}(\phi_{n+1}^{\tau+1})\|}{\|\boldsymbol{F}(\phi_{n+1}^{\tau+1})\|} \leq \epsilon_2 \tag{8.396}$$

For transient problems a good first guess ϕ_{n+1}^{0} is always available since the solution usually changes little between time steps, provided the time step length Δt_n is sufficiently small. All the more, if we use the error-controlled predictor-corrector methods as described in Sect. 8.13.5 an even better first guess can be obtained by using the predictor solution ϕ^p_{n+1} at the current time plane $n+1$, viz.,

$$\phi^0_{n+1} = \phi^p_{n+1} \tag{8.397}$$

where ϕ^p_{n+1} is given by (8.157) and (8.166) for the FE and AB scheme, respectively. Now, it is argued [211] that (1) the required degree of convergence is reached in just one iteration per time step when the predictor furnishes a sufficiently accurate first guess, and (2) the pre-set error measure ϵ used in the predictor-corrector schemes is recognized as the controlling parameter when keeping the time discretization error small. It leads to the so-called *one-step Newton method* (or alternatively, *one-step Picard method*), in which the predictor value ϕ^p_{n+1} is generally utilized to linearize the complete nonlinear system without any need for a repeated iteration within each time step. The following procedures result

$$\begin{aligned}
\boldsymbol{J}(\phi^p_{n+1}) \cdot \Delta\phi_{n+1} &= -\boldsymbol{R}(\phi^p_{n+1}) \\
\Delta\phi_{n+1} &= \phi_{n+1} - \phi^p_{n+1} \\
\boldsymbol{J}(\phi^p_{n+1}) &= \frac{\partial \boldsymbol{R}(\phi^p_{n+1})}{\partial \phi^p_{n+1}} \\
\boldsymbol{R}(\phi^p_{n+1}) &= \boldsymbol{A}(\phi^p_{n+1}) \cdot \phi^p_{n+1} - \boldsymbol{b}(\phi^p_{n+1})
\end{aligned} \tag{8.398}$$

for the one-step Newton method and

$$\boldsymbol{A}(\phi^p_{n+1}) \cdot \phi_{n+1} = \boldsymbol{b}(\phi^p_{n+1}) \tag{8.399}$$

for the one-step Picard method. The one-step Newton (or Picard) method embedded in the predictor-corrector scheme with its automatic step-size (time approximation error) control has shown a cost-effective and favorable approach in many applications. In comparison to the Picard method the extra work for the one-step Newton method is small in forming the Jacobian and residual matrices which can be done simultaneously with assembling the matrix system, however, in favor of achieving a quadratic convergence behavior. This is particularly true for an unsymmetric system matrix $\boldsymbol{A}$, typically appearing in ADE problems. For symmetric systems possessing strong nonlinearities the use of the Newton procedure can also be advantageous, in spite of losing symmetry in the final equation system to be solved.

8.19 Derived Quantities

8.19.1 Computing First Derivatives at Nodes

We have discussed in Sect. 3.11 the suitably chosen primary variables in form of hydraulic head, species concentration or temperature for solving the governing flow, mass and heat transport equations in porous media. Having known the

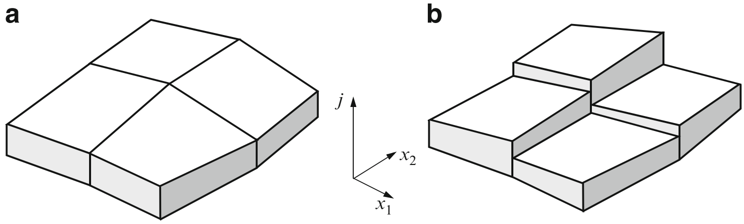


Fig. 8.47 (**a**) Continuous (smoothed) and (**b**) discontinuous (unsmoothed) flux component given for an element patch of quadrilaterals (Modified from [251])

spatio-temporal solutions of the primary variables, there is a need to obtain the solution of secondary variables, such as Darcy velocity, mass flux or heat flux, which represent quantities derived from the primary variables. In terms of the prototypical ADE equation (8.3) or (8.5), the FEM leads to the solution of the primary variable ϕ in space and time, which represents an elementwise continuous approximation (cf. Sect. 8.7). A derived quantity would be the flux

$$\boldsymbol{j} = -\boldsymbol{D} \cdot \nabla \phi \tag{8.400}$$

where $\boldsymbol{D}$ is a dispersion tensor. Since the finite element approximation of the primary variable ϕ is of the form (cf. (8.16))

$$\phi(\boldsymbol{x}, t) = \sum_j N_j(\boldsymbol{x})\, \phi_j(t) \tag{8.401}$$

we obtain the discrete flux

$$\boldsymbol{j}(\boldsymbol{x}, t) = -\sum_j \boldsymbol{D} \cdot \nabla N_j(\boldsymbol{x})\, \phi_j(t) \tag{8.402}$$

As result of the basic finite element solution, e.g., (8.322), ϕ_j is known at each global node j of the mesh and given time t so that $\boldsymbol{j}$ can be evaluated in a postprocessing operation. However, we recognize from (8.402) the first derivative in the flux $\boldsymbol{j}$ is no more continuous since the used element shape function N_j satisfies only C_0–continuity (cf. Sect. 8.7). Indeed, by using lower order elements of linear or quadratic type, the first derivatives do not possess anymore inter-element continuity. The elemental fluxes become discontinuous between elements and no unique fluxes at nodal points result as illustrated in Fig. 8.47.

Unfortunately, the discontinuity of the derived fluxes results in a number of serious drawbacks. Most important are balance errors arising in local flux evaluations. For example, Yeh [578] showed balance errors up to 30 % on a domain interior. On the other hand, the evaluation of streamlines according to (2.94) and pathlines according to (2.98) needs always a basically continuous flow field, otherwise coherent trajectories in the global flow field cannot be computed. Hence, suitable methods are required to produce continuous and precise fluxes over the finite element mesh, which are referred to as *smoothing* strategies. Most important are global and local smoothing as well as superconvergent patch recovery (SPR) techniques to derive continuous fluxes at internal nodes.

8.19.1.1 Global Smoothing

Global smoothing represents a natural approach of FEM to obtain continuous flux values at nodes. Most common is the technique basically proposed by Hinton and Campbell [251], which has proved to be quite widely used [590]. Yeh [578] firstly introduced such type of global smoothing in groundwater modeling to compute precise Darcy fluxes. A global finite element approximation of a smoothed (continuous) flux $\tilde{\boldsymbol{j}}$ can be written as

$$\tilde{\boldsymbol{j}}(\boldsymbol{x},t) = \sum_j N_j(\boldsymbol{x})\,\tilde{\boldsymbol{j}}_j(t) \tag{8.403}$$

Suppose an unsmoothed (discontinuous) flux is given by $\boldsymbol{j}$ (8.402), then the smooth function which provides a best fit in the least squares sense over the domain $\varOmega$ can be obtained from a minimization of the functional

$$\mathcal{I} = \int_{\varOmega} (\tilde{\boldsymbol{j}} - \boldsymbol{j})^2 d\varOmega \Rightarrow \text{min} \tag{8.404}$$

The minimization procedure

$$\begin{aligned} \frac{\partial \mathcal{I}}{\partial \tilde{\boldsymbol{j}}_i} &= \int_{\varOmega} 2(\tilde{\boldsymbol{j}} - \boldsymbol{j}) \frac{\partial \tilde{\boldsymbol{j}}}{\partial \tilde{\boldsymbol{j}}_i} d\varOmega = \boldsymbol{0} \\ &= \int_{\varOmega} N_i(\tilde{\boldsymbol{j}} - \boldsymbol{j}) d\varOmega \quad\;\; = \boldsymbol{0} \end{aligned} \qquad \text{for} \quad i = 1, 2, \ldots, N_{\mathrm{P}} \tag{8.405}$$

results in a system of linear equations to solve for the nodal vector of smoothed fluxes $\tilde{\boldsymbol{j}}_d$ for each vector component $d = 1, \ldots, D$ in $\Re^D$, viz.,

$$\boldsymbol{O} \cdot \tilde{\boldsymbol{j}}_d = \boldsymbol{F}_d \quad (d = 1, \ldots, D) \tag{8.406}$$

where $\boldsymbol{O}$ represents a mass (smoothing) matrix and $\boldsymbol{F}_d$ is the RHS $d-$component flux vector involving the unsmoothed relations. They are formed in the finite element assembling procedure as

$$\begin{aligned} \boldsymbol{O} &= O_{ij} = \sum_e \Big(\sum_I \sum_J O^e_{IJ} \Delta^e_{Ii} \Delta^e_{Jj} \Big) \\ \boldsymbol{F}_d &= F_{id} = \sum_e \Big(\sum_I F^e_{Id} \Delta^e_{Ii} \Big) \end{aligned} \tag{8.407}$$

where the element matrix and element vector are formed by

$$\begin{aligned} O^e_{IJ} &= \int_{\Omega^e} N^e_I N^e_J d\Omega^e \\ F^e_{Id} &= -\int_{\Omega^e} N^e_I \sum_J^{N_{\text{BN}}} \sum_l^D \big(D^e_{dl} \frac{\partial N^e_J}{\partial x_l} \phi^e_J\big) d\Omega^e \end{aligned} \qquad (d,l = 1,\ldots,D) \tag{8.408}$$

It can be easily seen that the least squares global smoothing is equivalent to the Galerkin weak statement directly applied to (8.400). The smoothing matrix $\boldsymbol{O}$ and the RHS vector $\boldsymbol{F}_d$ may be evaluated using numerical integration as described in Sect. 8.12. However, the complete solution of the linear matrix system (8.406) has been found too costly and furthermore in many cases unnecessary. The following smoothing procedures will be accordingly the preferred techniques.

A cost-effective alternative to (8.406) appears if the element smoothing matrix $\boldsymbol{O}$ is lumped by a row-summing technique (see Sect. 8.13.2) for each element e

$$O^e_{IJ} = \delta_{IJ} \int_{\Omega^e} N^e_I d\Omega^e \tag{8.409}$$

In doing so, there is no need anymore to solve the linear equation system (8.406). Instead, the smoothed flux can be explicitly evaluated by

$$\tilde{\boldsymbol{j}}_d = \boldsymbol{O}^{-1} \cdot \boldsymbol{F}_d \quad (d = 1,\ldots,D) \tag{8.410}$$

where the inverse of the diagonal matrix $\boldsymbol{O}^{-1}$ effects for each node a division by the sum $\sum_e \int_{\Omega^e} d\Omega^e$ of the surrounding element patch. This lumped form of global smoothing (8.410) can be recognized as an area/volume-weighted averaging for nodal flux values. However, this area/volume-weighing strategy has an essential drawback for irregular meshes: It weights larger elements more than smaller elements notwithstanding that larger elements imply presumably less accurate flux computations. To weight the more accurate smaller elements than the less accurate larger elements, the inverse area/volume-weighted averaging could be chosen instead [209], however, its foundation lies outside of the Galerkin-FEM framework. Thus, the following local smoothing and recovery strategies will be preferred.

8.19.1.2 Superconvergent Flux Evaluation and Local Smoothing

In FEM there is the phenomenon of *superconvergence* [590], which is referred to optimal sampling points for which derivatives are more accurate than elsewhere. In particular, Gauss quadrature sampling points (cf. Sect. 8.12) exhibit superconvergent behavior and have shown the suited locations $\boldsymbol{x}$ to evaluate derived quantities

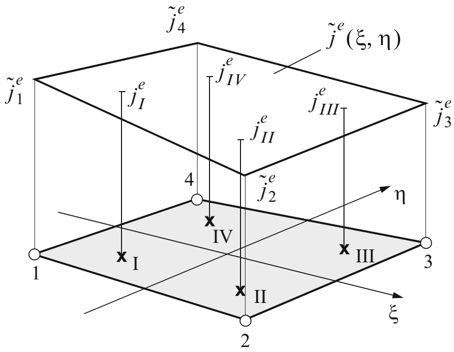


Fig. 8.48 Superconvergent flux components $j_I^e, j_{II}^e, j_{III}^e, j_{IV}^e$ sampled at Gauss points ξ_p, η_p $(p = 1, \ldots, 4)$ and functional dependency of smoothed flux $\tilde{j}^e(\xi, \eta)$ to extrapolate to nodal flux components $\tilde{j}_1^e, \tilde{j}_2^e, \tilde{j}_3^e, \tilde{j}_4^e$ for a linear quadrilateral element e

having best accuracy.[17] Such a superconvergent flux evaluation means that the discontinuous flux $\boldsymbol{j}(\boldsymbol{x}(\boldsymbol{\eta}), t)$ of (8.402) has to be sampled at the Gauss points with the local coordinates $\boldsymbol{\eta}_p$ $(p = 1, \ldots, m)$ within each element e, i.e.,

$$\boldsymbol{j}(\boldsymbol{x}(\boldsymbol{\eta}), t) \rightarrow \boldsymbol{j}^e(\boldsymbol{\eta}_p, t) = -\sum_J^{N_{\mathrm{BN}}} \boldsymbol{D}^e \cdot \nabla N_J^e(\boldsymbol{\eta}_p)\, \phi_J^e(t) \quad (p = 1, \ldots, m) \tag{8.411}$$

where $\boldsymbol{\eta}_p$ is the vector of local coordinates, e.g., (ξ_p, η_p, ζ_p) for a 3D element, and m is the total number of Gauss points. Their locations in 2D and 3D elements are displayed in Fig. 8.17. Conveniently, m is chosen by the same number of element nodes N_{BN} so that the Gauss sample points can be related to corresponding element nodes (see Fig. 8.48).

Now, the smoothing of the discontinuous flux is considered over individual elements, termed *local smoothing*. It is assumed that the smoothed flux function $\tilde{\boldsymbol{j}}^e(\boldsymbol{\eta})$ is a least squares fit to the selected values $\boldsymbol{j}^e(\boldsymbol{\eta}_p)$ at the Gauss points $p = 1, \ldots, m$ for each element e separately. In using the least squares procedure of Sect. 8.19.1.1 to an individual element the following local equation system results

$$\begin{pmatrix} \int_{\Omega^e} N_1^e N_1^e d\Omega^e & \ldots & \int_{\Omega^e} N_1^e N_{N_{\mathrm{BN}}}^e d\Omega^e \\ \int_{\Omega^e} N_2^e N_1^e d\Omega^e & \ldots & \int_{\Omega^e} N_2^e N_{N_{\mathrm{BN}}}^e d\Omega^e \\ \vdots & \vdots & \vdots \\ \int_{\Omega^e} N_{N_{\mathrm{BN}}}^e N_1^e d\Omega^e & \ldots & \int_{\Omega^e} N_{N_{\mathrm{BN}}}^e N_{N_{\mathrm{BN}}}^e d\Omega^e \end{pmatrix} \cdot \begin{pmatrix} \tilde{j}_{1d}^e \\ \tilde{j}_{2d}^e \\ \vdots \\ \tilde{j}_{N_{\mathrm{BN}}d}^e \end{pmatrix} = \begin{pmatrix} \int_{\Omega^e} N_1^e j_d^e(\boldsymbol{\eta}_1) d\Omega^e \\ \int_{\Omega^e} N_2^e j_d^e(\boldsymbol{\eta}_2) d\Omega^e \\ \vdots \\ \int_{\Omega^e} N_{N_{\mathrm{BN}}}^e j_d^e(\boldsymbol{\eta}_{m=N_{\mathrm{BN}}}) d\Omega^e \end{pmatrix}$$

$$(d = 1, \ldots, D) \tag{8.412}$$

[17] Superconvergence of the derivatives can be shown for the Gauss points, at least for quadrilateral elements [590]. On the other hand, the location of the superconvergent points for triangular elements is not fully known. Zienkiewicz and Zhu [594] propose to use optimal points, for instance the central points for linear triangles.

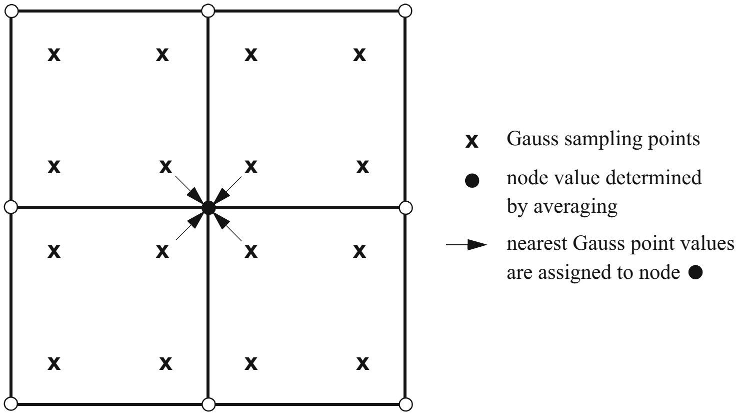

x Gauss sampling points
● node value determined by averaging
→ nearest Gauss point values are assigned to node ●

Fig. 8.49 Element patch surrounding the particular node ● at which a globally smoothed flux is computed on the basis of extrapolated or shifted superconvergent flux values sampled at Gauss points ×

to solve the smoothed flux $d-$components $\tilde{j}^e_{Id}$ at the local nodes $I = 1, 2, \ldots, N_{\text{BN}}$ of an element e from the superconvergent flux $d-$components $j^e_d(\boldsymbol{\eta}_p)$ according to (8.411) sampled at Gauss points $p = 1, \ldots, m = N_{\text{BN}}$. Simple relations are obtained from (8.412) for elements having a constant Jacobian in $d\Omega^e = |\boldsymbol{J}^e| d\boldsymbol{\eta}$, see derivations in Appendix H. For example, for a rectangular and parallelogram 2D element by using 2 × 2 Gauss points (listed in Table 8.2) the following expression can be derived

$$\begin{pmatrix} \tilde{j}^e_{1d} \\ \tilde{j}^e_{2d} \\ \tilde{j}^e_{3d} \\ \tilde{j}^e_{4d} \end{pmatrix} = \begin{pmatrix} 1+\frac{\sqrt{3}}{2} & -\frac{1}{2} & 1-\frac{\sqrt{3}}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1+\frac{\sqrt{3}}{2} & -\frac{1}{2} & 1-\frac{\sqrt{3}}{2} \\ 1-\frac{\sqrt{3}}{2} & -\frac{1}{2} & 1+\frac{\sqrt{3}}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1-\frac{\sqrt{3}}{2} & -\frac{1}{2} & 1+\frac{\sqrt{3}}{2} \end{pmatrix} \cdot \begin{pmatrix} j^e_{dI} \\ j^e_{dII} \\ j^e_{dIII} \\ j^e_{dIV} \end{pmatrix} \qquad (d = 1, \ldots, D) \tag{8.413}$$

to compute directly the smoothed flux components $\tilde{j}^e_{Id}$ at the corner nodes $I = 1, 2, 3, 4$ from the superconvergent flux components at Gauss points I, II, III, IV (illustrated in Fig. 8.48), where $j^e_{dI} = j^e_d(-\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}})$, $j^e_{dII} = j^e_d(\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}})$, $j^e_{dIII} = j^e_d(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$ and $j^e_{dIV} = j^e_d(-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$. It can be easily seen that such a formula of local smoothing represents nothing more than a local scheme to interpolate/extrapolate Gauss point values to nodal point values [251].

Unlike global smoothing, the local smoothing strategy does not produce unique flux values at nodes and therefore an appropriate averaging of the superconvergent flux values is needed. Consider for example the element patch surrounding the particular node at which a unique flux has to be computed as shown in Fig. 8.49. For each element of the patch the superconvergent flux values determined at the Gauss sampling points can be either (1) interpolated/extrapolated to the particular node by using solutions of (8.412), exemplified for 2D rectangular elements in form of (8.413), or (2) without any interpolation, by a simple assignment (shift) of Gauss-point fluxes nearest to the particular node. The latter strategy is commonly acceptable, in particular for linear elements, because the derivatives usually vary only slightly or are even constant within the element. In doing so, each nodal contribution of elements sharing the particular node is summed up and finally averaged in the following ways. The simplest method is the arithmetic mean for the global node i in the form

$$\tilde{\boldsymbol{j}}_i = \frac{1}{N_\Sigma} \sum_e^{N_\Sigma} \boldsymbol{j}_i^e \tag{8.414}$$

to determine the smoothed flux $\tilde{\boldsymbol{j}}_i$ at the node i, where N_Σ is the number of patch elements surrounding the node i and $\boldsymbol{j}_i^e$ is the superconvergent flux of element e assigned or extrapolated to node i. Alternatively, as discussed above, an inverse area/volume-weighted averaging can be favorable to attain an improved approximation for more irregularly shaped elements of a patch. It reads

$$\tilde{\boldsymbol{j}}_i = \frac{1}{\sum_e^{N_\Sigma} w^e} \sum_e^{N_\Sigma} w^e \boldsymbol{j}_i^e \tag{8.415}$$

with the weights

$$w^e = \begin{cases} \frac{1}{\Delta x^e} & \text{1D} \\ \frac{1}{A^e} & \text{2D} \\ \frac{1}{V^e} & \text{3D} \end{cases} \tag{8.416}$$

The averaging techniques in combination with local smoothing the superconvergent flux values have proved to be accurate comparable to global smoothing procedures. Instead of nodal averaging, however, an improved method exists in which a polynomial expansion is used on an element patch fitting locally the superconvergent points in a least squares manner, known as superconvergent patch recovery (SPR) to be described next.

8.19.1.3 Superconvergent Patch Recovery (SPR)

Zienkiewicz and Zhu [594] have proposed a powerful and accurate method of computing derivatives via a direct polynomial smoothing, which leads to superconvergent flux values at *all* and not only at certain Gauss sampling points within the finite element, called *superconvergent patch recovery* (SPR). In this method a polynomial expansion of the function $\tilde{\boldsymbol{j}}(\boldsymbol{x})$ describing the derivatives is used on an element patch surrounding the interelement node at which the nodal derivatives

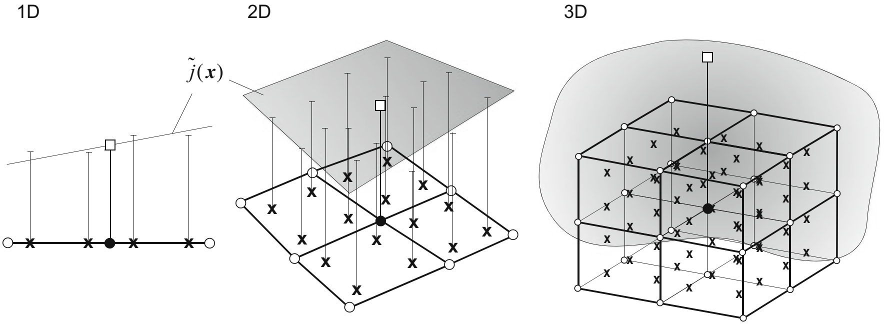


Fig. 8.50 Element patches of linear quadrilateral elements in 1D, 2D and 3D. Polynomial expansion of linear function $\tilde{j}(\boldsymbol{x})$ describing the derivatives at superconvergent Gauss points × is used on the element patch surrounding the node ● at which recovery is desired. Its nodal value □ is obtained by evaluating the resulting polynomial

have to be determined (recovered). The polynomial is chosen in the same order as occurring in the used finite element approximation of the primary variable, which achieves superconvergent accuracy everywhere if this polynomial is made to fit the superconvergent Gauss sampling points in a least square manner [209, 590]. Let us consider for example the linear polynomial expansion of the derivatives applied to element patches of linear quadrilateral elements in 1D, 2D and 3D as shown in Fig. 8.50. The following three working steps are needed:

1. The derivatives are evaluated at the superconvergent Gauss points ×.
2. A least-squares fit through the Gauss points is made with a linear polynomial.
3. The superconvergent nodal derivatives □ are obtained by evaluating the resulting polynomial at the patch node ●.

Having determined the derivatives $\boldsymbol{j}(\boldsymbol{x})$ at the Gauss points according to (8.402) we introduce the linear polynomials for the superconvergent (smooth) derivatives in the form

$$\tilde{\boldsymbol{j}}(\boldsymbol{x}) = \begin{cases} \alpha + \beta\, x & \text{1D} \\ \alpha + \beta\, x + \chi\, y + \delta\, xy & \text{2D} \\ \alpha + \beta\, x + \chi\, y + \delta\, z + \epsilon\, xy + \phi\, yz + \gamma\, zx + \eta\, xyz & \text{3D} \end{cases} \tag{8.417}$$

where $\boldsymbol{x} = (x \;\; y \;\; z)^T$ are the Cartesian coordinates and $\alpha, \beta, \chi, \delta, \epsilon, \phi, \gamma, \eta$ are unknown coefficients to be determined. Note that the mixed terms in (8.417) does not exist for 2D triangular and 3D tetrahedral elements. Now, the method of least-squares is applied to minimize the sum of the squares $\tilde{\boldsymbol{j}}(\boldsymbol{x}_i) - \boldsymbol{j}(\boldsymbol{x}_i)$ over all Gauss points $i = 1, 2, \ldots, n$ encountered in the element patch, i.e.,

$$\mathcal{I} = \tfrac{1}{2} \sum_{i=1}^{n} \big[\tilde{\boldsymbol{j}}(\boldsymbol{x}_i) - \boldsymbol{j}(\boldsymbol{x}_i)\big]^2 \Rightarrow \min \tag{8.418}$$

where $n = mN_\Sigma$ (m = number of Gauss points per element, N_Σ = number of patch elements). For example, the minimization of $\mathcal{I} = \frac{1}{2}\sum_{i=1}^{n}[\alpha + \beta\, x_i + \chi\, y_i + \delta\, x_i\, y_i - \boldsymbol{j}(x_i, y_i)]^2$ for the 2D polynomial with respect to the four unknown coefficients yields

$$\begin{aligned}
\partial\mathcal{I}/\partial\alpha &= 0 = \textstyle\sum_i \alpha + \beta\, x_i + \chi\, y_i + \delta\, x_i\, y_i - \boldsymbol{j}(x_i, y_i)\\
\partial\mathcal{I}/\partial\beta &= 0 = \textstyle\sum_i [\alpha + \beta\, x_i + \chi\, y_i + \delta\, x_i\, y_i - \boldsymbol{j}(x_i, y_i)]x_i\\
\partial\mathcal{I}/\partial\chi &= 0 = \textstyle\sum_i [\alpha + \beta\, x_i + \chi\, y_i + \delta\, x_i\, y_i - \boldsymbol{j}(x_i, y_i)]y_i\\
\partial\mathcal{I}/\partial\delta &= 0 = \textstyle\sum_i [\alpha + \beta\, x_i + \chi\, y_i + \delta\, x_i\, y_i - \boldsymbol{j}(x_i, y_i)]x_i\, y_i
\end{aligned} \tag{8.419}$$

This leads to the local 4×4 linear system ($\sum \equiv \sum_{i=1}^{n}$)

$$\begin{pmatrix}
n & \sum x_i & \sum y_i & \sum x_i y_i\\
\sum x_i & \sum x_i^2 & \sum x_i y_i & \sum x_i^2 y_i\\
\sum y_i & \sum x_i y_i & \sum y_i^2 & \sum x_i y_i^2\\
\sum x_i y_i & \sum x_i^2 y_i & \sum x_i y_i^2 & \sum x_i^2 y_i^2
\end{pmatrix} \cdot \begin{pmatrix}\alpha\\ \beta\\ \chi\\ \delta\end{pmatrix} = \begin{pmatrix}
\sum \boldsymbol{j}(x_i, y_i)\\
\sum x_i \boldsymbol{j}(x_i, y_i)\\
\sum y_i \boldsymbol{j}(x_i, y_i)\\
\sum x_i y_i \boldsymbol{j}(x_i, y_i)
\end{pmatrix} \tag{8.420}$$

to solve for the polynomial coefficients α, β, χ and δ. Then, the recovered derivative at an interelement node j can be easily computed from

$$\tilde{\boldsymbol{j}}(x_j, y_j) = \alpha + \beta\, x_j + \chi\, y_j + \delta\, x_j\, y_j \tag{8.421}$$

Similar recovery expressions can be derived for 1D and 3D element patches. The additional numerical cost in SPR is acceptable because the equation system like (8.420) remains small. The total effort usually is smaller than global smoothing and larger than local smoothing, however, in favor of an improved accuracy of the derivatives at the nodes. A robust implementation of SPR requires that the rank of the resulting local equation system, e.g., (8.420), must be equivalent to the number of terms a used in the polynomial expansion [326]:

$$n \geq a \tag{8.422}$$

where $a = 2$ in 1D, $a = 4$ for quadrilateral and $a = 3$ for triangular elements in 2D and $a = 8$ for quadrilateral and $a = 4$ for tetrahedral elements in 3D. Thus, there is a minimal number of elements N_Σ in an element patch to make the resulting local equation system solvable. Hence, for linear triangles N_Σ has to be greater than or equal to three. To overcome this difficulty in a robust recovery procedure the number of sampling points m is set at least equal to the number of terms a in the polynomial expansion regardless of achieving actual superconvergence for the recovered solution.

8.19.2 Computing First Derivatives at Exterior or Interior Boundaries: The Consistent Boundary Flux Method (CBFM) and Budget Analysis

8.19.2.1 Consistently Derived Boundary Flux Based on Weak Forms

In the previous Sect. 8.19.1 appropriate postprocessing methods for determining fluxes $\boldsymbol{j} = -\sum_j \boldsymbol{D}\cdot\nabla N_j\,\phi_j$ at nodal points are introduced. Now, we could assume that those nodal fluxes are also suitable to evaluate *boundary* fluxes q_n in a way such as

$$q_n = -\sum_j \phi_j\,(\boldsymbol{D}\cdot\nabla N_j)\cdot\boldsymbol{n}\big|_\Gamma \tag{8.423}$$

where $\boldsymbol{n}$ is the unit normal vector to the boundary Γ and ϕ_j is the given solution of the primary variable at nodal points j. Boundary fluxes are needed for evaluating balance quantities in a budget analysis, for example balanced boundary fluxes through exterior Dirichlet-type boundary Γ_D or Cauchy-type boundary section Γ_C of the model domain Ω or through interior boundaries Γ_I of subdomains Ω_I as part of Ω (Fig. 8.51). However, the numerical differentiation in the form (8.423) is not a sufficiently accurate and reasonable expression of a discrete boundary flux because it does not guarantee a proper balance condition at the local position of the boundary. Additionally, (8.423) requires an actual construction of $\boldsymbol{n}$, which is cumbersome and often quite ambiguous if the boundary is not smooth. In a sum, the discrete boundary flux in the form of (8.423) has not the required quality of a locally balanced flux and is accordingly rather inappropriate for any balance evaluation.

To overcome the difficulties with (8.423) the *consistent boundary flux method* (CBFM) satisfies the requirements for local balance accuracy as suggested by Gresho et al. [213]. It has been shown that CBFM (and related methods) leads to conservative (consistent) flux quantities, e.g., [47, 69, 148, 355, 403]. To obtain a consistent approximation to the boundary flux

$$q_n = \begin{cases} (\phi\boldsymbol{q} - \boldsymbol{D}\cdot\nabla\phi)\cdot\boldsymbol{n}\big|_\Gamma & \text{for the divergence form of ADE} \\ -\boldsymbol{D}\cdot\nabla\phi\cdot\boldsymbol{n}\big|_\Gamma & \text{for the convective form of ADE} \end{cases} \tag{8.424}$$

we directly utilize the weak statements (8.46) and (8.53) of the governing balance equations for divergence form and convective form, respectively. Applying the Galerkin finite element weighting we find the appropriate weak formulation

$$\begin{aligned}\int_\Gamma N_i\,q_n\,d\Gamma = -\int_\Omega N_i\frac{\partial(\mathcal{R}\phi)}{\partial t}d\Omega + \int_\Omega \phi\boldsymbol{q}\cdot\nabla N_i d\Omega -\\ \int_\Omega \nabla N_i\cdot(\boldsymbol{D}\cdot\nabla\phi)d\Omega - \int_\Omega N_i(\vartheta\phi - H - Q_{\phi w})d\Omega\end{aligned} \tag{8.425}$$

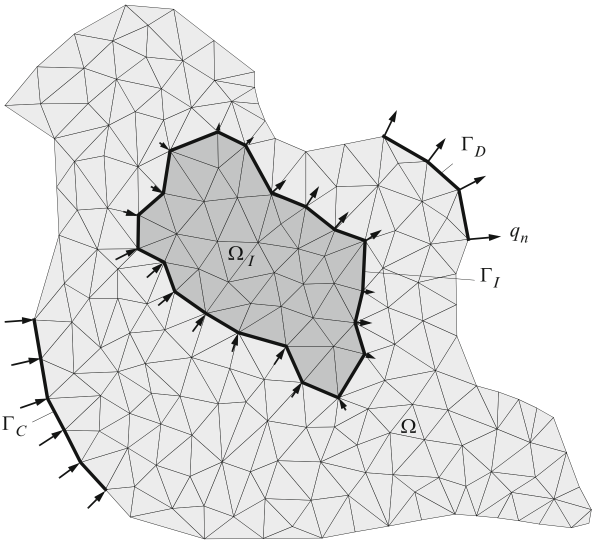


Fig. 8.51 Mesh of domain $\varOmega$ with discrete exterior boundary sections $\varGamma_D$ and $\varGamma_C$ as well as interior boundary $\varGamma_I$ enclosing subdomain $\varOmega_I \subset \varOmega$

for the divergence form of ADE and

$$\int_{\varGamma} N_i \, q_n \, d\varGamma = -\int_{\varOmega} N_i \acute{\mathcal{R}} \frac{\partial \phi}{\partial t} d\varOmega - \int_{\varOmega} N_i \boldsymbol{q} \cdot \nabla \phi d\varOmega - \\ \int_{\varOmega} \nabla N_i \cdot (\boldsymbol{D} \cdot \nabla \phi) d\varOmega - \int_{\varOmega} N_i [(\vartheta + Q)\phi - H - Q_{\phi w}] d\varOmega \qquad (8.426)$$

for the convective form of ADE, where the primary variable

$$\phi = \sum_j N_j \phi_j \qquad (8.427)$$

is now known from the approximate finite element solution ϕ_j given at each nodal point j and current time t_{n+1}. These weak formulations allow a consistent computation of the boundary flux q_n. In doing so, we expand q_n in the finite element context as

$$q_n = \sum_j N_j q_{nj} \qquad (8.428)$$

where q_{nj} is the nodal boundary flux to be determined on $\varGamma$ and at evaluation time t_{n+1}. Inserting (8.427) and (8.428) into (8.425) and (8.426) the following matrix system results

$$\boldsymbol{M} \cdot \boldsymbol{q}_n = -\boldsymbol{O} \cdot \dot{\boldsymbol{\phi}} - (\boldsymbol{A} + \boldsymbol{C} + \boldsymbol{R}) \cdot \boldsymbol{\phi} + \boldsymbol{Q} \qquad (8.429)$$

where

$$\boldsymbol{q}_n = q_{nj} = \begin{pmatrix} q_{n1} \\ q_{n2} \\ \vdots \\ q_{nN_{\mathrm{P}}} \end{pmatrix} \tag{8.430}$$

is the nodal vector of the boundary flux and

$$\begin{aligned} \boldsymbol{M} &= M_{ij} = \sum_e \Big(\sum_I \sum_J M_{IJ}^e \Delta_{Ii}^e \Delta_{Jj}^e \Big) \\ M_{IJ}^e &= \int_{\Gamma^e} N_I^e N_J^e d\Gamma^e \end{aligned} \tag{8.431}$$

is the boundary mass matrix, which couples q_{nj} to its nearest neighbors of Γ. The matrices $\boldsymbol{O}$, $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{R}$ as well as the RHS vector $\boldsymbol{Q}$ appearing in (8.429) are already given by (8.103)–(8.105). The assembly of (8.429) is done in the usual way at element level, except that only those elements with nodes on Γ need be considered, i.e., the linear matrix sytem (8.429) is solved only for a subset of N_{P} nodes because all contributions to nodes which do not belong to Γ are irrelevant. The linear system (8.429) is solved for the nodal boundary flux $\boldsymbol{q}_n$ on Γ, where the RHS of (8.429) is built up with the known solution ϕ and its time derivative $\dot{\phi}$ at evaluation time t_{n+1}. We recognize that the CBFM is a strategy in which the 'forward' solution system (8.100) is reversely solved on $\Gamma-$nodes with known ϕ and $\dot{\phi}$. Babuška and Miller [18] have shown that the consistent boundary fluxes exhibit superior convergence behavior, i.e., superconvergence.

Remark. The equation (8.429) represents the *consistently derived* flux having the following remarkable properties: (1) If this flux is computed on a Dirichlet boundary Γ_D, it will lead to the same ϕ when imposed as a Neumann-type BC, i.e., $\boldsymbol{q}_n$ and ϕ are equivalent and exchangeable as BC's. This means that with a known ϕ the domain Ω can arbitrarily be subdivided into subdomains $\Omega_I \subset \Omega$ (Fig. 8.51) forming nonoverlapping interior and/or exterior boundaries of Dirichlet type (ϕ is prescribed there) formed along mesh edges/faces Γ_I on which the consistent boundary flux is computable. (2) The boundary flux guarantees the appropriate approximation to the governing balance equation both globally and locally. The smallest subdomain can be even each single element $\Omega_I \rightarrow \Omega^e$ so that the boundary flux on $\Gamma_I \rightarrow \Gamma^e$ also guarantees conservation according to the local balance with (8.429), see Sect. 8.19.3 for further discussion.

8.19.2.2 Lumped Solution

To avoid the solution of the linear system (8.429) the cost-effective alternative is to invoke mass lumping for $\boldsymbol{M}$, i.e.,

$$\boldsymbol{M} = M_{ij} = \delta_{ij} \int_{\Gamma} N_i \, d\Gamma \tag{8.432}$$

With mass lumping (8.432) the nodal boundary fluxes in (8.429) become uncoupled and can be explicitly computed from

$$\boldsymbol{q}_n = -\boldsymbol{M}^{-1} \cdot \left[\boldsymbol{O} \cdot \dot{\boldsymbol{\phi}} + (\boldsymbol{A} + \boldsymbol{C} + \boldsymbol{R}) \cdot \boldsymbol{\phi} - \boldsymbol{Q}\right] \tag{8.433}$$

8.19.2.3 Integral Boundary Flux

Alternatively to (8.433), we can simply sum up the contributions of the system for each row i to obtain the *integral* boundary balance flux at the boundary node i

$$\begin{aligned} Q_{ni} &= -\int_{\Gamma} N_i \, q_n d\Gamma \\ &= -\sum_j M_{ij} q_{nj} \\ &= \sum_j \left[O_{ij} \dot{\phi}_j + (A_{ij} + C_{ij} + R_{ij}) \phi_j - Q_i\right], \quad (j = 1, \ldots, N_{\mathrm{P}}) \end{aligned} \tag{8.434}$$

or in matrix form

$$\begin{aligned} \boldsymbol{Q}_n &= -\boldsymbol{M} \cdot \boldsymbol{q}_n \\ &= \boldsymbol{O} \cdot \dot{\boldsymbol{\phi}} + (\boldsymbol{A} + \boldsymbol{C} + \boldsymbol{R}) \cdot \boldsymbol{\phi} - \boldsymbol{Q} \end{aligned} \tag{8.435}$$

where the sign of $Q_{ni} = \boldsymbol{Q}_n$ is used in accordance with the definitions of well-type SPC terms (cf. Sect. 6.3), i.e., a positive $\boldsymbol{Q}_n$ corresponds to a point sink. The integral boundary flux $\boldsymbol{Q}_n = -\boldsymbol{M} \cdot \boldsymbol{q}_n$ is used in a *budget analysis* in which the balance quantities on boundaries Γ are determined at evaluation time t_{n+1}. It is also required in constraint formulations for BC's (see Sect. 6.4).

8.19.2.4 Auxiliary Problem Formulation for Convective Form of ADE

In use of the convective form of ADE the boundary flux is dispersion/diffusion-controlled $q_n^d = -\boldsymbol{D} \cdot \nabla \phi \cdot \boldsymbol{n}\big|_{\Gamma}$ (8.424) according to the basic weak statement. For a budget analysis it is also desired to quantify the missing advective part of a boundary flux $q_n^a = \phi \boldsymbol{q} \cdot \boldsymbol{n}\big|_{\Gamma}$, where $\boldsymbol{q}$ is the advective flux. We recall that the convective form of ADE results from the substitution of mass conservation, cf. (3.45). To obtain the total boundary flux $q_n = q_n^a + q_n^d$ we have to retrieve the substituted mass conservation via an auxiliary weak formulation. Let us consider for example the mass conservation equation given in Table 3.7 and multiplying all terms by ϕ. It results

$$\phi\Big(sS_o\frac{\partial h}{\partial t}+\varepsilon\frac{\partial s}{\partial t}\Big)+\phi\nabla\cdot\boldsymbol{q}=\phi(Q+Q_{\mathrm{EOB}}) \tag{8.436}$$

Its weak statement reads

$$\int_\Omega w\phi\nabla\cdot\boldsymbol{q}\,d\Omega=\int_\Omega w\phi(Q+Q_{\mathrm{EOB}})d\Omega-\int_\Omega w\phi\Big(sS_o\frac{\partial h}{\partial t}+\varepsilon\frac{\partial s}{\partial t}\Big)d\Omega \tag{8.437}$$

where h, the hydraulic head, and s, the saturation, are another primary variables which are assumed to be known from a separate finite element solution of the flow equation. Using the product rule of differentiation

$$\nabla\cdot(w\phi\boldsymbol{q})=\phi\boldsymbol{q}\cdot\nabla w+w\phi\nabla\cdot\boldsymbol{q}+w\boldsymbol{q}\cdot\nabla\phi \tag{8.438}$$

and employing the Gauss's integral theorem (2.77) on the LHS term of (8.438) we obtain from (8.437)

$$\begin{aligned}\int_\Gamma w\phi\boldsymbol{q}\cdot\boldsymbol{n}\,d\Gamma=&\int_\Omega \phi\nabla w\cdot\boldsymbol{q}\,d\Omega+\int_\Omega w\nabla\phi\cdot\boldsymbol{q}\,d\Omega+\\&\int_\Omega w\phi(Q+Q_{\mathrm{EOB}})d\Omega-\int_\Omega w\phi\Big(sS_o\frac{\partial h}{\partial t}+\varepsilon\frac{\partial s}{\partial t}\Big)d\Omega\end{aligned} \tag{8.439}$$

Now, using the Galerkin weak formulation $w\rightarrow N_i$, invoking the Darcy law to express the flow vector as $\boldsymbol{q}=-k_r\boldsymbol{K}f_\mu\cdot(\nabla h+\chi\boldsymbol{e})$ (cf. Table 3.7) and expanding the known variables $\phi=\sum_j N_j\phi_j$, $h=\sum_j N_jh_j$ and $s=\sum_j N_js_j$ in the finite element context, we find

$$\begin{aligned}\int_\Gamma N_iq_n^a d\Gamma=&-\sum_j\int_\Omega \nabla N_i\cdot[k_r\boldsymbol{K}f_\mu\cdot(\nabla N_j+\chi\boldsymbol{e})]h_j\big(\sum_l N_l\phi_l\big)d\Omega-\\&\sum_j\int_\Omega N_i\nabla N_j\,\phi_j\cdot\big[\sum_l k_r\boldsymbol{K}f_\mu\cdot(\nabla N_l\,h_l+\chi\boldsymbol{e})\big]d\Omega+\\&\int_\Omega N_i\big(\sum_l N_l\phi_l\big)(Q+Q_{\mathrm{EOB}})d\Omega-\\&\int_\Omega N_i\big(\sum_l N_l\phi_l\big)\Big[\big(\sum_l N_ls_l\big)S_o\big(\sum_l N_l\frac{\partial h_l}{\partial t}\big)+\varepsilon\big(\sum_l N_l\frac{\partial s_l}{\partial t}\big)\Big]d\Omega\end{aligned} \tag{8.440}$$

or with expanding $q_n^a=\sum_j N_jq_{nj}^a$

$$\boldsymbol{M}\cdot\boldsymbol{q}_n^a=-\boldsymbol{U}(\boldsymbol{\phi})\cdot\boldsymbol{h}-\boldsymbol{V}(\boldsymbol{h})\cdot\boldsymbol{\phi}+\boldsymbol{X}(\boldsymbol{\phi})-\boldsymbol{Y}(\boldsymbol{\phi},\boldsymbol{s},\dot{\boldsymbol{h}},\dot{\boldsymbol{s}}) \tag{8.441}$$

to solve the advective boundary flux vector $\boldsymbol{q}_n^a$ at evaluation time t_{n+1}, where the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are related to the 1st and 2nd RHS-terms and the vectors $\boldsymbol{X}$

and $\boldsymbol{Y}$ are related to the 3rd and 4th RHS-terms of (8.440). It is assumed that the solutions $\boldsymbol{\phi}$, $\boldsymbol{h}$, $\boldsymbol{s}$ and the time derivatives $\dot{\boldsymbol{h}}$ and $\dot{\boldsymbol{s}}$ are known. Dependence of the solution vectors $\boldsymbol{\phi}$, $\boldsymbol{h}$, $\boldsymbol{s}$, $\dot{\boldsymbol{h}}$ and/or $\dot{\boldsymbol{s}}$ in $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are shown in parentheses. Finally, we can combine (8.429) and (8.441) to find the expression for solving the total consistent boundary flux $\boldsymbol{q}_n = \boldsymbol{q}_n^a + \boldsymbol{q}_n^d$ in the form

$$\begin{aligned}\boldsymbol{M} \cdot \boldsymbol{q}_n = -\boldsymbol{O} \cdot \dot{\boldsymbol{\phi}} - [\boldsymbol{A} + \boldsymbol{C} + \boldsymbol{R} + \boldsymbol{V}(\boldsymbol{h})] \cdot \boldsymbol{\phi} - \boldsymbol{U}(\boldsymbol{\phi}) \cdot \boldsymbol{h} +& \\ \boldsymbol{X}(\boldsymbol{\phi}) - \boldsymbol{Y}(\boldsymbol{\phi}, \boldsymbol{s}, \dot{\boldsymbol{h}}, \dot{\boldsymbol{s}}) + \boldsymbol{Q}&\end{aligned} \tag{8.442}$$

associated with the convective form of ADE.

8.19.2.5 Illustrative Example

To clarify the consistent flux method let us consider a simple, however, quite representative and illustrative example [213]: A steady-state diffusion problem with a varying source in one dimension x. The corresponding basic PDE is

$$-\nabla^2\phi = H(x), \quad 0 \le x \le 3 \tag{8.443}$$

which has to be solved for $\phi = \phi(x)$ subject to the BC's

$$\phi = \begin{cases} 0 & \text{at} \quad x = 0 \quad \text{and} \\ 0 & \text{at} \quad x = 3 \end{cases} \tag{8.444}$$

and with the source function

$$H(x) = \begin{cases} 0 & \text{for} \quad 0 \le x < 2 \quad \text{and} \\ 6 & \text{for} \quad 2 \le x \le 3 \end{cases} \tag{8.445}$$

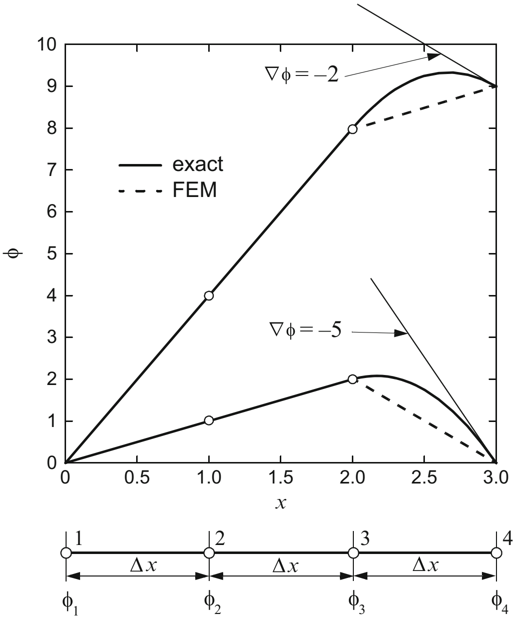
The exact solution is

$$\phi(x) = \begin{cases} x & \text{for} \quad 0 \le x \le 2 \\ -3x^2 + 13x - 12 & \text{for} \quad 2 \le x \le 3 \end{cases} \tag{8.446}$$

which is plotted as the lower solid curve in Fig. 8.52. The *exact* boundary flux $q_n = -\nabla\phi \cdot \boldsymbol{n} = -\partial\phi/\partial x|_\Gamma$ through the outer boundary at $x = 3$ can be simply derived from (8.446) as $q_n = 5$.

The problem is approximated by using just three linear elements, each of unit length $\Delta x = 1$ (Fig. 8.52). The finite element discretization leads to the following matrix system (see Appendix H.1, note that incoming and outgoing gradients are canceled at interior element boundaries due to their opposite signs):

Fig. 8.52 Steady diffusion problem in one dimension



$$\frac{1}{\Delta x}\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix} = \begin{pmatrix} q_n \\ 0 \\ H\frac{\Delta x}{2} \\ H\frac{\Delta x}{2} - q_n \end{pmatrix} \tag{8.447}$$

The BC's (8.444) giving $\phi_1 = \phi_4 = 0$ are incorporated in (8.447), cf. Sect. 8.16. Then, the following discrete equations result

$$\tfrac{1}{\Delta x}(2\phi_2 - \phi_3) = 0 \tag{8.448}$$

and

$$\tfrac{1}{\Delta x}(-\phi_2 + 2\phi_3) = H\tfrac{\Delta x}{2} \tag{8.449}$$

Hence, with $\phi_2 = 1$ and $\phi_3 = 2$ the finite element solution is exact at the nodes (see the dashed line in the lower curve of Fig. 8.52).

Now, suppose that the flux q_n at the boundary $x = 3$ is desired. If the conventional nodal flux evaluation according to (8.423) is employed, we find (cf. Appendix H.1)

$$q_n = -\tfrac{1}{\Delta x}(-\phi_3 + \phi_4) = 2 \tag{8.450}$$

which is different to the exact solution of 5. In contrast to (8.450), the consistent boundary flux results from (8.447) (resolving the last row for q_n with the given $\phi-$solution):

$$q_n = H\frac{\Delta x}{2} - \frac{1}{\Delta x}(-\phi_3 + \phi_4) = 5 \tag{8.451}$$

which agrees with the exact solution. It is obvious that although (8.450) is in fact the true slope of the approximate solution, the CBFM solution (8.451) yields the correct balanced flux, which properly accounts for both the source term in the finite element and the diffusion at $x = 3$. We can even proof the local balance for each element if we solve the consistent boundary flux for a separate element. Consider, for example, the last element e, $(2 \leq x \leq 3)$:

$$\frac{1}{\Delta x^e}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \phi_3 \\ \phi_4 \end{pmatrix} = \begin{pmatrix} H^e\frac{\Delta x^e}{2} \\ H^e\frac{\Delta x^e}{2} \end{pmatrix} - \begin{pmatrix} -q_n^- \\ q_n^+ \end{pmatrix} \tag{8.452}$$

It yields $q_n^- = -1$ and $q_n^+ = 5$ as left-sided and right-sided boundary flux of the element, respectively. Thus, the element balance is exactly satisfied with

$$H^e\Delta x^e + q_n^- - q_n^+ = 6 - 1 - 5 = 0 \tag{8.453}$$

Finally, to demonstrate consistency in the finite element solutions, let us solve the problem by using a Neumann-type BC q_N at $x = 3$, where we use the derived value for q_n from (8.450): $q_N = q_n = 2$. The corresponding nodal equations are for this case

$$\begin{aligned} \frac{1}{\Delta x}(2\phi_2 - \phi_3) &= 0 \\ \frac{1}{\Delta x}(-\phi_2 + 2\phi_3 - \phi_4) &= H\frac{\Delta x}{2} \\ \frac{1}{\Delta x}(-\phi_3 + \phi_4) &= H\frac{\Delta x}{2} - q_N \end{aligned} \tag{8.454}$$

The solution to (8.454) for $q_N = 2$ is $\phi_2 = 4$, $\phi_3 = 8$ and $\phi_4 = 9$, which is displayed as the dashed upper curve in Fig. 8.52 in comparison to the exact solution for $q_N = 2$ given as $\phi = 4x$ for $0 \leq x \leq 2$ and $\phi = -3x^2 + 16x - 12$ for $2 \leq x \leq 3$. On the other hand, using $q_N = 5$ from (8.451) we retrieve the original result as $\phi_2 = 2$, $\phi_3 = 3$ and $\phi_4 = 0$. Thus, it demonstrates that only the consistently derived flux can be applied as a natural BC to recover the original solution obtained with Dirichlet-type BC's. Although we have shown only a 1D problem, the same essential issues are given in multidimensional and transient problems [209, 213, 277].

8.19.3 Continuous Finite Element Approach Is Locally Conservative

The basic model equations which are solved via approximate methods represent balance laws for conserving physical quantities such as mass, momentum and energy.

Accordingly, the used numerical approach should also respect these conservation equations both globally and locally. *Conservativity* (see definition in Sect. 1.2.2) enforces that incoming and outgoing fluxes through interior and exterior boundaries of a global domain and its subdivided subdomains have to be conserved and consistent with the source/sink and storage effects occurring in the balance volumes, otherwise the method is nonconservative which can produce artificial sources and sinks, changing the balance both locally and globally. Nonconservative methods are to be declined to avoid erroneous solutions.

Local conservativity means that conservation is guaranteed for each of the smallest discrete unit, i.e., for each element (or cell), regardless of mesh (grid) size. However, local conservativity does not mean local accuracy. The problem solution may be inaccurate, but will, nevertheless, be conservative. Repeatedly, it is believed that finite element methods are not locally conservative. But, this is a misbelief (and in part a strange discrediting of FEM) which has been refuted in a number of papers, see e.g., [47, 89, 148, 277, 355]. Obviously, there is a misunderstanding on both the basic conservation law structure of the FEM and the computation of local fluxes. A major reason is apparently in the misuse of nodal derivatives in form of (8.402) or (8.423) as balance fluxes. Indeed, those nonconsistent fluxes obtained from a numerical differentiation are not necessarily conservative and can cause significant local balance errors [47, 148, 355, 578]. The simple example of Sect. 8.19.2.5 has evidently shown the importance of a suitable flux computation for balance evaluations.

The present continuous finite element approach is based on an elementwise *continuous* approximation, see Sect. 8.7. It guarantees continuity up to first derivatives (fluxes) even at element interfaces. Having this property for C_0 continuous basis functions, the subdivision of the global integrals into subdomains, elements and subboundaries can be done via (8.62) without any interelement residual. As a consequence, fluxes between adjacent elements cancel since the flux is contained within the element, while fluxes exposed on the external boundary do not. With other word, the fluxes appears only on external (global) boundaries, while fluxes interchanging between adjacent elements remain hidden during the usual computation. However, we can evaluate this type of flux as consistent boundary flux q_n. In Sect. 8.19.2 the CBFM is described which provides precise boundary fluxes at any exterior or interior boundaries of a meshed domain coinciding with the element edges or faces. If the external boundary $\varGamma$ is used, the consistent boundary fluxes determine the global conservation of the domain $\bar{\varOmega} = \varOmega \cup \varGamma$, if the boundary $\varGamma_I$ refers to a subdomain $\varOmega_I \subset \varOmega$, such as illustrated in Fig. 8.51, the boundary flux measures the exchange between the adjacent subdomain and accordingly determines the conservation of the subdomain $\bar{\varOmega}_I = \varOmega_I \cup \varGamma_I$, and finally if the boundary is even chosen as the element boundary $\varGamma^e$, the resulting boundary flux measures the conservation of the single element $\bar{\varOmega}^e = \varOmega^e \cup \varGamma^e$ (Fig. 8.53).

Now, we can utilize the weak formulations (8.425) and (8.426) to find the boundary flux q_n^e of each single element e in form of *element conservation laws*:

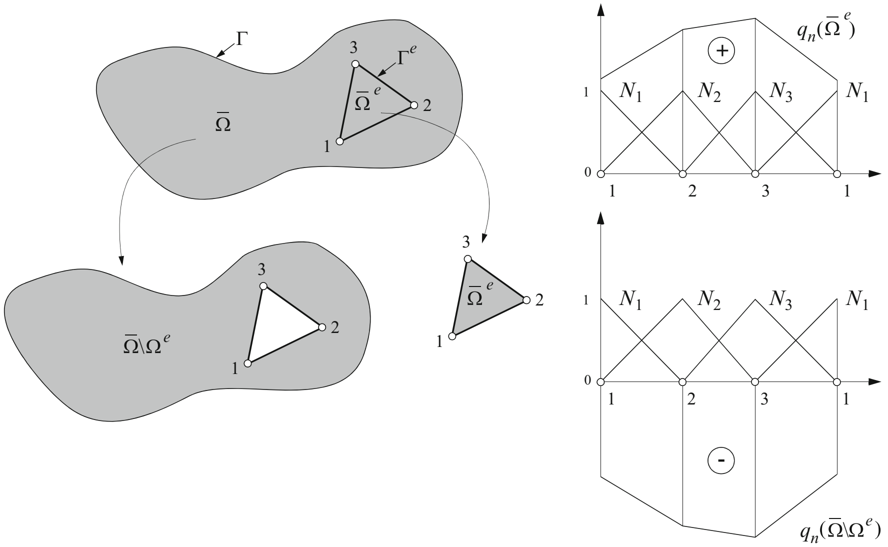


Fig. 8.53 Consistent boundary flux $q_n(\bar{\Omega}^e)$ of element $\bar{\Omega}^e$ in equilibrium with boundary flux $q_n(\bar{\Omega}\backslash\Omega^e)$ of subdomain $\bar{\Omega}\backslash\Omega^e$ (Modified from [277])

$$\int_{\Gamma^e} N_I^e\, q_n^e\, d\Gamma^e = -\int_{\Omega^e} N_I^e \frac{\partial(\mathcal{R}^e\phi^e)}{\partial t} d\Omega^e + \int_{\Omega^e} \phi^e \boldsymbol{q}^e \cdot \nabla N_I^e d\Omega^e - \int_{\Omega^e} \nabla N_I^e \cdot (\boldsymbol{D}^e \cdot \nabla\phi^e) d\Omega^e - \int_{\Omega} N_I^e(\vartheta^e\phi^e - H^e - Q_{\phi w}^e) d\Omega^e \tag{8.455}$$

for the divergence form of ADE and

$$\int_{\Gamma^e} N_I^e\, q_n^e\, d\Gamma^e = -\int_{\Omega^e} N_I^e \acute{\mathcal{R}}^e \frac{\partial\phi^e}{\partial t} d\Omega^e - \int_{\Omega^e} N_I^e \boldsymbol{q}^e \cdot \nabla\phi^e d\Omega^e - \int_{\Omega^e} \nabla N_I^e \cdot (\boldsymbol{D}^e \cdot \nabla\phi^e) d\Omega^e - \int_{\Omega^e} N_I^e[(\vartheta^e + Q^e)\phi^e - H^e - Q_{\phi w}^e] d\Omega^e \tag{8.456}$$

for the convective form of ADE, where $I = 1, \ldots, N_{\mathrm{BN}}$. Similar to (8.434) and (8.435), respectively, we can summarize (8.455) and (8.456) as follows

$$\begin{aligned} Q_{nI}^e &= -\int_{\Gamma^e} N_I^e\, q_n^e d\Gamma^e \\ &= -\sum_J M_{IJ}^e\, q_{nJ}^e \\ &= \sum_J \big[O_{IJ}\dot{\phi}_J^e + (A_{IJ}^e + C_{IJ}^e + R_{IJ}^e)\phi_J^e - Q_I\big], \quad (J = 1, \ldots, N_{\mathrm{BN}}) \end{aligned} \tag{8.457}$$

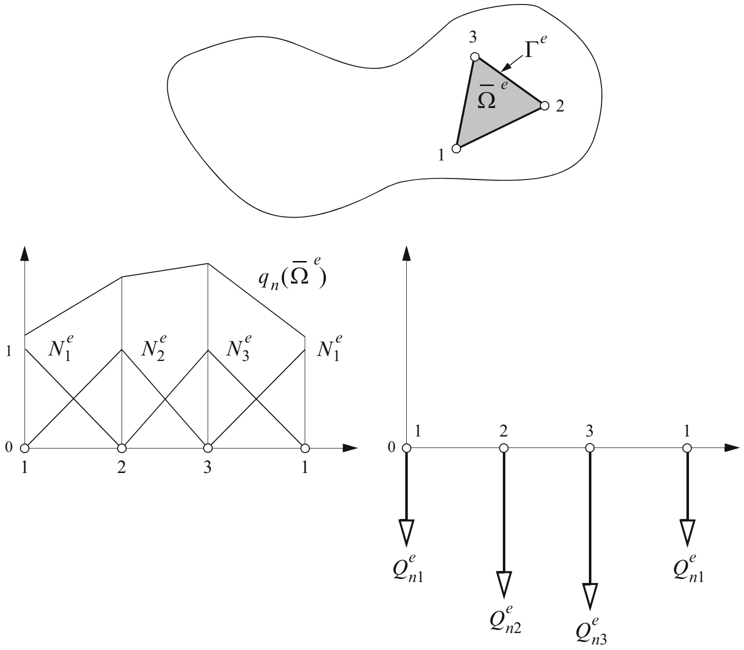


Fig. 8.54 Element conservation: The consistent boundary flux $q_n(\bar{\Omega}^e)$ is the conservative redistribution of the integral element flux Q_{nI}^e in terms of the element basis functions N_I^e, $I = 1, \ldots, N_{\text{BN}}$ (Modified from [277])

or in matrix form

$$\begin{aligned} \boldsymbol{Q}_n^e &= -\boldsymbol{M}^e \cdot \boldsymbol{q}_n^e \\ &= \boldsymbol{O}^e \cdot \dot{\boldsymbol{\phi}}^e + (\boldsymbol{A}^e + \boldsymbol{C}^e + \boldsymbol{R}^e) \cdot \boldsymbol{\phi}^e - \boldsymbol{Q}^e \end{aligned} \tag{8.458}$$

with

$$M_{IJ}^e = \int_{\Gamma^e} N_I^e N_J^e d\Gamma^e \tag{8.459}$$

where $\boldsymbol{q}_n^e$ is the element boundary flux and $\boldsymbol{Q}_n^e = \boldsymbol{O}^e \cdot \dot{\boldsymbol{\phi}}^e + (\boldsymbol{A}^e + \boldsymbol{C}^e + \boldsymbol{R}^e) \cdot \boldsymbol{\phi}^e - \boldsymbol{Q}^e$ is the integral element flux (Fig. 8.54). By summing (8.457) over $I = 1, \ldots, N_{\text{BN}}$, we see that

$$\int_{\Gamma^e} q_n^e d\Gamma^e + \sum_I Q_{nI}^e = 0 \tag{8.460}$$

which represents the element conservation. Thus, the sum of the integral element fluxes is a conserved quantity. The corresponding element boundary flux q_n^e of element $\bar{\Omega}^e$ is in equilibrium with the boundary fluxes of the adjacent complementary subdomain $\bar{\Omega} \backslash \Omega^e$ (see Fig. 8.53), viz.,

$$q_n(\bar{\Omega}^e) = -q_n(\bar{\Omega}\backslash\Omega^e) \tag{8.461}$$

It becomes clear that the nodal fluxes Q^e_{nI} and their continuous redistribution q^e_n in terms of the element basis functions are different but equivalent representations of the same information [277], viz.,

$$\sum_J \int_{\Gamma^e} N^e_I N^e_J q^e_{nJ} d\Gamma^e = -Q^e_{nI} \tag{8.462}$$

$$Q^e_{nI} = -\int_{\Gamma^e} N^e_I\, q^e_n d\Gamma^e \tag{8.463}$$

Once Q^e_{nI} is known, $q^e_n = \sum_J N^e_J q^e_{nJ}$ is uniquely defined by (8.462). Likewise, if $q^e_n = q_n(\bar{\Omega}^e)$ is known, the nodal fluxes Q^e_{nI} are uniquely defined by (8.463). These quantities are fundamental to the local conservativity of the continuous FEM.

8.19.4 Note on Mixed Finite Element Formulations

So far we have considered the basic balance equation in a form in which the governing flux $\boldsymbol{j} = -\boldsymbol{D}\cdot\nabla\phi$ has been suitably substituted so that only one unknown function ϕ, the primary variable, remains in the scalar governing equations (8.3) or (8.5). This elimination of $\boldsymbol{j}$ leads to a mathematically well-defined problem with appropriate BC's expressed in terms of ϕ or its gradients (8.4) or (8.6). In the FEM context it leads to an approximation of only one unknown variable ϕ_i per node i of a mesh, i.e., degrees of freedom are $N_{\text{DOF}} = 1$, and the resulting matrix system becomes usually easily solvable. However, in this approach the required knowledge of the secondary variable in form of the flux $\boldsymbol{j}$ must be obtained as a derived quantity, which naturally implies a loss of accuracy compared to the accuracy attainable for the primary variable, notwithstanding the precise evaluation techniques for deriving $\boldsymbol{j}$ such as described in the preceding Sect. 8.19.1.

The FEM does not restrict per se the formulation to governing equations in which the flux is eliminated. It is also possible to refer to a formulation where both ϕ and $\boldsymbol{j}$ are chosen as primary variables. This is called as a *mixed finite element formulation*, e.g., [56, 84, 436, 590]. Mixed finite element methods are inevitable in CFD for solving the coupled system of Navier-Stokes equations [209], where the eliminatation of fluxes (velocities) from the basic equations is not possible or restricted. This is quite different to Darcy-based flow equations in porous media, where a mixed formulation appears as a useful but commonly nonessential alternative [75, 152, 378]. To illustrate the mixed finite element formulation for the present class of problems, let us write the governing ADE (8.5) in the alternative form as

$$\begin{aligned} \acute{\mathcal{R}}\frac{\partial \phi}{\partial t} + \boldsymbol{q}\cdot\nabla\phi + \nabla\cdot\boldsymbol{j} + (\vartheta + Q)\phi &= H + Q_{\phi w} \\ \boldsymbol{j} &= -\boldsymbol{D}\cdot\nabla\phi \end{aligned} \tag{8.464}$$

and introduce the finite element approximation for both primary variables ϕ and $\boldsymbol{j}$ as (cf. (8.16))

$$\begin{aligned} \phi(\boldsymbol{x},t) &= \textstyle\sum_j N_j(\boldsymbol{x})\,\phi_j(t) \\ \boldsymbol{j}(\boldsymbol{x},t) &= \textstyle\sum_l M_d(\boldsymbol{x})\,\boldsymbol{j}_d(t) \end{aligned} \tag{8.465}$$

where N_j and M_d represent basis functions at global nodes j and d, respectively, which must not coincide, and ϕ_j and $\boldsymbol{j}_d$ are the corresponding nodal vectors of the unknowns ϕ and $\boldsymbol{j}$, respectively. Now, taking (8.465) and applying the GFEM to (8.464), in which the weighting functions are in accordance with the basis functions, we can find the following matrix system

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{C}^{\dagger} & \boldsymbol{0} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\phi} \\ \boldsymbol{J} \end{pmatrix} = \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{0} \end{pmatrix} \tag{8.466}$$

to solve simultaneously $\boldsymbol{\phi} = \phi_j$ and $\boldsymbol{J} = \boldsymbol{j}_d$, where

$$\begin{aligned} \boldsymbol{A} = A_{ij} &= \sum_e \Big(\int_{\Omega^e} \acute{\mathcal{R}}^e\, N_i N_j \frac{\partial}{\partial t} d\Omega^e + \int_{\Omega^e} N_i \boldsymbol{q}^e \cdot \nabla N_j \, d\Omega^e + \\ &\quad \int_{\Omega^e} (\vartheta^e + Q^e) N_i N_j \, d\Omega^e + \int_{\Gamma_C^e} \Phi^e N_i N_j \, d\Gamma^e \Big) - \\ &\quad \delta_{ij} Q_w(t)\big|_i \\ \boldsymbol{C} = C_{id} &= \sum_e \int_{\Omega^e} N_i \nabla M_d \, d\Omega^e \\ \boldsymbol{C}^{\dagger} = C_{lj}^{\dagger} &= -\sum_e \int_{\Omega^e} M_l \boldsymbol{D} \cdot \nabla N_j \, d\Omega^e \\ \boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i H^e d\Omega^e + \int_{\Gamma_C^e} N_i \Phi^e \phi_C^e \, d\Gamma^e - \int_{\Gamma_N^e} N_i q_N^e \, d\Gamma^e \Big) - \\ &\quad \phi_w Q_w(t)\big|_i \end{aligned} \tag{8.467}$$

in which the indices i, j and l, d, respectively, run over the same nodal points. The mixed finite element formulation in the form of (8.467) includes the following properties:

1. The simultaneous solution of $\boldsymbol{\phi}$ and $\boldsymbol{J}$ leads to a higher accuracy of the flux $\boldsymbol{J}$ compared to the standard formulation (at the same mesh resolution) in which $\boldsymbol{J}$ has been eliminated and appears as secondary variable. Indeed, $\boldsymbol{J}$ resulting from the mixed formulation satisfies implicitly local conservativity.

2. The higher accuracy of $\boldsymbol{J}$ must be paid by a significant increase in the computational effort because the increased degrees of freedom (e.g., $N_{\mathrm{DOF}} = 4$ in 3D) considerably enlarge the final equation system to be solved.
3. The resulting matrix system (8.466) forms a saddle point problem, where the total matrix is not positive definite. It can lead to difficulties in the solving the equations.
4. The formulation of BC's for the flux is restricted.
5. The mixed interpolation for ϕ and $\boldsymbol{j}$ must satisfy a compatibility condition, known as LBB (Ladyshenkaya-Babuška-Brezzi) condition, see e.g., [56, 149, 209], otherwise the mixed formulation does not guarantee stability. The basis functions N_j and M_d are differently chosen. Once subjected to 1st-order derivatives they have to be C_0 continuous functions, otherwise no continuity is needed. A well-known stable element is the Taylor-Hood element [209], in which the flux $\boldsymbol{j}$ is interpolated quadratically and the scalar variable ϕ is interpolated by a linear continuous function such as used by Diersch [130] in free convection flow modeling in porous media. Other useful stable elements are discussed in [30, 56, 149, 209, 436], where the Raviart-Thomas element [75] appears suitable for the present class of porous-media problems [31, 46, 359, 476].

The mixed finite element formulation offered the possibility for obtaining a potentially higher accuracy in the flux computations compared to a standard formulation (simulated at the same mesh resolution), however, at a significant increase in the computational effort and at the expense of a reduced robustness and flexibility. This limits the mixed FEM to only specific (often academic, small-size) problems. In practical modeling of flow and transport processes in porous and fractured media, mixed finite element formulations are neither feasible nor necessary in general. Indeed, the accuracy of fluxes achieved from the standard formulations by using the derived quantity evaluations as discussed in Sect. 8.19.1 are usually able to provide equivalently precise flux computations. In the following, we need not to resort to mixed finite element formulations.

Chapter 9
Flow in Saturated Porous Media: Groundwater Flow

9.1 Introduction

The numerical simulation of the groundwater flow is one of the standard tasks for modelers in the field of subsurface hydrology. Groundwater refers to subsurface water, where the porous medium is *fully saturated*. Typically, groundwater modeling is concerned with the motion of subsurface water in aquifers and aquifer systems, which can be unconfined or confined (see definitions in Sect. 2.2.1), i.e., bounded by free surface(s) or without the presence of those. Solutions are required for fully 3D, vertical or essentially horizontal 2D and axisymmetric *isothermal* flow of *homogeneous* flow in saturated porous media with and without free surface(s). Variably saturated porous media and variable-density flow will be subject of Chaps. 10 and 11, respectively.

9.2 Basic Equations

9.2.1 3D, Vertical 2D and Axisymmetric Problems

The basic equations have been developed in Sect. 3.10.6 and can be taken from Table 3.9 for isothermal flow of homogeneous fluids (i.e., no density coupling) in saturated porous media. It yields

$$\begin{aligned} S_o \tfrac{\partial h}{\partial t} + \nabla \cdot \boldsymbol{q} &= Q \\ \boldsymbol{q} &= -\boldsymbol{K} \cdot \nabla h \end{aligned} \tag{9.1}$$

to be solved for the hydraulic head h and the Darcy velocity $\boldsymbol{q}$, where S_o is the specific storage coefficient, $\boldsymbol{K}$ is the tensor of hydraulic conductivity and Q is

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_9,

a general source/sink function (Table 3.9). Usually, $\boldsymbol{q}$ is substituted by the Darcy equation to obtain the governing Richards-type equation (cf. Sect. 3.11) in the form:

$$S_o \frac{\partial h}{\partial t} - \nabla \cdot (\boldsymbol{K} \cdot \nabla h) = Q_h + Q_{hw} \tag{9.2}$$

where the source/sink term $Q = Q_h + Q_{hw}$ is suitably split into a supply term Q_h and a well-type SPC term Q_{hw}. The PDE (9.2) has to be solved for the remaining primary variable h subject to a set of BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC as introduced in Sect. 6.3.1,

$$\begin{aligned} h &= h_D && \text{on} \quad \Gamma_D \times t[t_0, \infty) \\ -(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} &= q_h && \text{on} \quad \Gamma_N \times t[t_0, \infty) \\ -(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} &= -\Phi_h (h_C - h) && \text{on} \quad \Gamma_C \times t[t_0, \infty) \\ Q_{hw} &= -\textstyle\sum_w Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty) \end{aligned} \tag{9.3}$$

also the nonlinear BC of a free surface as introduced in Sect. 6.5.1,

$$\left.\begin{aligned} -(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} &= \varepsilon_e \frac{\partial h}{\partial t} - P \\ h &= z \end{aligned}\right\} \quad \text{on} \quad \Gamma_S \times t[t_0, \infty) \tag{9.4}$$

and in combination with the IC of the form

$$h(\boldsymbol{x}, t_0) = h_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\Omega} \tag{9.5}$$

where ε_e is the specific yield, z corresponds to the vertical coordinate (aligned to the gravity direction) and the total boundary is $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_C \cup \Gamma_S$. Once the hydraulic head has been solved, the secondary variable of Darcy velocity $\boldsymbol{q} = -\boldsymbol{K} \cdot \nabla h$ can be evaluated as a derived quantity of known h. The essential parameters required for solving (9.2) with (9.3)–(9.5) are listed in Tables I.1–I.6 of Appendix I in accordance with the chosen problem type. *Steady-state* flow situations occur if $S_o = 0$ (and $\varepsilon_e = 0$ for $\Gamma_S \neq \emptyset$) or $\partial h / \partial t$ approaches to zero.[1]

9.2.2 Horizontal 2D Flow in Unconfined Aquifers

The basic equations for the essentially horizontal, vertically averaged flow in unconfined aquifers have been developed in Sect. 3.10.7 and summarized in Table 3.10. We find

[1] Optionally, FEFLOW suppresses the time derivative term $\partial h / \partial t$ for solving steady-state solutions.

$$\begin{aligned} (BS_o + \varepsilon_e)\tfrac{\partial h}{\partial t} + \nabla \cdot \bar{\boldsymbol{q}} &= \bar{Q} \\ \bar{\boldsymbol{q}} &= -B\,\boldsymbol{K} \cdot \nabla h \end{aligned} \tag{9.6}$$

with the variably discharging aquifer thickness

$$B = h - f^B \tag{9.7}$$

to be solved for the hydraulic head h and the depth-integrated Darcy velocity $\bar{\boldsymbol{q}} = B\boldsymbol{q}$, where $S = BS_o + \varepsilon_e$ appears as an effective storage coefficient and f^B is the bottom bounding surface of aquifer (Table 3.10). Substituting the Darcy velocity $\bar{\boldsymbol{q}}$ in the mass conservation equation we obtain the following Richards-type equation valid for essentially horizontal 2D flow in unconfined aquifers:

$$(BS_o + \varepsilon_e)\frac{\partial h}{\partial t} - \nabla \cdot (B\,\boldsymbol{K} \cdot \nabla h) = \bar{Q}_h + \bar{Q}_{hw} \tag{9.8}$$

where the source/sink term $\bar{Q} = \bar{Q}_h + \bar{Q}_{hw}$ is suitably split into a depth-integrated supply term $\bar{Q}_h$ and a depth-integrated well-type SPC term $\bar{Q}_{hw}$. The solution of (9.8) for the primary variable h is associated with the following BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC (cf. Sect. 6.3.1)[2]

$$\begin{aligned} h &= h_D && \text{on} \quad \varGamma_D \times t[t_0, \infty) \\ -(B\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} &= Bq_h && \text{on} \quad \varGamma_N \times t[t_0, \infty) \\ -(B\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} &= -B\varPhi_h(h_C - h) && \text{on} \quad \varGamma_C \times t[t_0, \infty) \\ \bar{Q}_{hw} &= -\textstyle\sum_w Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \varOmega \times t[t_0, \infty) \end{aligned} \tag{9.9}$$

imposed on $\varGamma = \varGamma_D \cup \varGamma_N \cup \varGamma_C$ and with the IC of the form

$$h(\boldsymbol{x}, t_0) = h_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{9.10}$$

The secondary variable of the Darcy velocity $\bar{\boldsymbol{q}} = -B\boldsymbol{K} \cdot \nabla h$ is evaluated in a postprocessing computation as a derived quantity of known h. The essential parameters required for solving (9.8) with (9.9) and (9.10) are listed in Table I.8 of Appendix I.

[2] Special Neumann-type and Cauchy-type BC-formulations exist for *integral* BC's (cf. Sect. 6.5.4):

$$\begin{aligned} -(B\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} &= \bar{q}_h && \text{on} \quad \varGamma_N \times t[t_0, \infty) \\ -(B\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} &= -\bar{\varPhi}_h(h_C - h) && \text{on} \quad \varGamma_C \times t[t_0, \infty) \end{aligned}$$

which incorporate the depth integration in the BC values $\bar{q}_h$ and $\bar{\varPhi}_h$. This can be beneficial to prevent a $h-$dependency in the flux BC's due to the variable aquifer thickness $B = h - f^B$. For example, a solution with prescribed Neumann influx q_h can lead to a consecutively reduced h and accordingly a reduced B so that the effective influx Bq_n at the boundary section unavoidably tends to zero. In contrast to q_h, a prescription of $\bar{q}_h$ would not feature such a self-reinforcing dependency.

9.2.3 Horizontal 2D Flow in Confined Aquifers

The basic equations for the essentially horizontal, vertically averaged flow in confined aquifers have been developed in Sect. 3.10.7 and summarized in Table 3.11. It holds

$$\begin{aligned} \bar{S}_o \tfrac{\partial h}{\partial t} + \nabla \cdot \bar{\boldsymbol{q}} &= \bar{Q} \\ \bar{\boldsymbol{q}} &= -\boldsymbol{T} \cdot \nabla h \end{aligned} \tag{9.11}$$

to be solved for the hydraulic head h and the depth-integrated Darcy velocity $\bar{\boldsymbol{q}}$, where $\bar{S}_o = BS_o$ is the depth-integrated specific storage coefficient and $\boldsymbol{T}$ is the tensor of transmissivity (3.302). Substituting the Darcy velocity $\bar{\boldsymbol{q}}$ in the mass conservation equation we obtain the following Richards-type equation valid for essentially horizontal 2D flow in confined aquifers:

$$\bar{S}_o \frac{\partial h}{\partial t} - \nabla \cdot (\boldsymbol{T} \cdot \nabla h) = \bar{Q}_h + \bar{Q}_{hw} \tag{9.12}$$

where the source/sink term $\bar{Q} = \bar{Q}_h + \bar{Q}_{hw}$ is suitably split into a depth-integrated supply term $\bar{Q}_h$ and a depth-integrated well-type SPC term $\bar{Q}_{hw}$. The solution of (9.12) for the primary variable h is associated with the following BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC (cf. Sect. 6.3.1)

$$\begin{aligned} h &= h_D && \text{on} \quad \varGamma_D \times t[t_0, \infty) \\ -(\boldsymbol{T} \cdot \nabla h) \cdot \boldsymbol{n} &= \bar{q}_h && \text{on} \quad \varGamma_N \times t[t_0, \infty) \\ -(\boldsymbol{T} \cdot \nabla h) \cdot \boldsymbol{n} &= -\bar{\varPhi}_h (h_C - h) && \text{on} \quad \varGamma_C \times t[t_0, \infty) \\ \bar{Q}_{hw} &= -\textstyle\sum_w Q_w(t) \delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \varOmega \times t[t_0, \infty) \end{aligned} \tag{9.13}$$

imposed on $\varGamma = \varGamma_D \cup \varGamma_N \cup \varGamma_C$ and with the IC of the form

$$h(\boldsymbol{x}, t_0) = h_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{9.14}$$

The secondary variable of the Darcy velocity $\bar{\boldsymbol{q}} = -\boldsymbol{T} \cdot \nabla h$ is evaluated in a postprocessing computation as a derived quantity of known h. The essential parameters required for solving (9.12) with (9.13) and (9.14) are listed in Table I.7 of Appendix I.

9.3 Finite Element Formulation

The fundamental concepts of FEM are thoroughly described in Chap. 8. Based on the given principles we use now the GFEM to solve the governing flow equations (9.2), (9.8) and (9.12) associated with the corresponding BC's and IC's for the different classes of flow problems stated above. For convenience we only develop in detail the finite element equations for the fully 3D, vertical 2D and axisymmetric problems of Sect. 9.2.1. The remaining formulations for the horizontal 2D flow equations in unconfined and confined aquifers will appear rather similar and can be easily deduced from the given developments.

9.3.1 Weak Form

According to Sect. 8.5 the weak form for (9.2) appears as a special case of the ADE weak statement deduced from the expression (8.53). We find

$$\int_{\Omega} w S_o \frac{\partial h}{\partial t} d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{K} \cdot \nabla h) d\Omega - \int_{\Omega} w (Q_h + Q_{hw}) d\Omega - \int_{\Gamma} w (\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n}\, d\Gamma = 0, \quad \forall w \in H^1(\Omega) \tag{9.15}$$

where w is a suitable weighting function. Separating the boundary integral of (9.15) into the four segments $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_C \cup \Gamma_S$ imposed by the Dirichlet, Neumann, Cauchy-type and free-surface BC's, respectively, we invoke the BC's and SPC of (9.3) and BC of (9.4) to obtain

$$\int_{\Omega} w S_o \frac{\partial h}{\partial t} d\Omega + \int_{\Gamma_S} w \varepsilon_e \frac{\partial h}{\partial t} d\Gamma + \int_{\Omega} \nabla w \cdot (\boldsymbol{K} \cdot \nabla h) d\Omega - \int_{\Omega} w Q_h d\Omega + \sum_{w} w(\boldsymbol{x}_w) Q_w(t) + \int_{\Gamma_N} w q_h d\Gamma - \int_{\Gamma_C} w \Phi_h (h_C - h) d\Gamma - \int_{\Gamma_S} w P d\Gamma = 0, \quad \forall w \in H_0^1(\Omega) \tag{9.16}$$

9.3.2 GFEM and Resulting Matrix Systems

Choosing the approximate functional form for the solution h

$$h(\boldsymbol{x}, t) \approx \sum_j N_j(\boldsymbol{x})\, h_j(t), \quad j = 1, \dots, N_{\mathrm{P}} \tag{9.17}$$

and using the Galerkin method with the weighting function

$$w \to w_i = N_i, \quad i = 1, \dots, N_{\mathrm{P}} \tag{9.18}$$

we find the following Galerkin-based finite element formulation of (9.16), viz.,

$$
\begin{aligned}
&\sum_e \int_{\Omega^e} N_i S_o^e \frac{\partial}{\partial t}(\sum_j N_j h_j) d\Omega^e + \sum_e \int_{\Gamma_S^e} N_i \varepsilon_e^e \frac{\partial}{\partial t}(\sum_j N_j h_j) d\Gamma^e + \\
&\sum_e \int_{\Omega^e} \nabla N_i \cdot \big[\boldsymbol{K}^e \cdot \nabla(\sum_j N_j h_j)\big] d\Omega^e - \sum_e \int_{\Omega^e} N_i Q_h^e d\Omega^e + Q_w(t)\big|_i + \\
&\sum_e \int_{\Gamma_N^e} N_i q_h^e d\Gamma^e - \sum_e \int_{\Gamma_C^e} N_i \Phi_h^e \big[h_C^e - (\sum_j N_j h_j)\big] d\Gamma^e - \sum_e \int_{\Gamma_S^e} N_i P^e d\Gamma^e = 0 \\
&\qquad 1 \le i, j \le N_{\mathrm{P}}
\end{aligned}
\tag{9.19}
$$

The assembly process is used to form the global matrix system of the spatial finite element discretization

$$
\boldsymbol{O} \cdot \dot{\boldsymbol{h}} + \boldsymbol{C} \cdot \boldsymbol{h} - \boldsymbol{F} = \boldsymbol{0} \tag{9.20}
$$

where

$$
\boldsymbol{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_{N_{\mathrm{P}}} \end{pmatrix}, \quad \dot{\boldsymbol{h}} = \begin{pmatrix} \frac{dh_1}{dt} \\ \frac{dh_2}{dt} \\ \vdots \\ \frac{dh_{N_{\mathrm{P}}}}{dt} \end{pmatrix} \tag{9.21}
$$

and the matrices and RHS vector

$$
\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \Big(\int_{\Omega^e} S_o^e N_i N_j d\Omega^e + \int_{\Gamma_S^e} \varepsilon_e^e N_i N_j d\Gamma^e \Big) \\
\boldsymbol{C} = C_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{K}^e \cdot \nabla N_j) d\Omega^e + \int_{\Gamma_C^e} \Phi_h^e N_i N_j d\Gamma^e \Big) \\
\boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i Q_h^e d\Omega^e + \int_{\Gamma_C^e} N_i \Phi_h^e h_C^e d\Gamma^e - \int_{\Gamma_N^e} N_i q_h^e d\Gamma^e + \\
&\qquad \int_{\Gamma_S^e} N_i P^e d\Gamma^e \Big) - Q_w(t)\big|_i
\end{aligned}
\tag{9.22}
$$

for $(i, j = 1, \ldots, N_{\mathrm{P}})$ and $(e = 1, \ldots, N_{\mathrm{E}})$. The integrals appearing in (9.22) are integrated on element level in the local coordinates (see Sect. 8.12). Analytical evaluations of partial integral terms of (9.22) can be deduced from developments done in Appendix H for selected element types. The differential elements $d\Omega^e$ and $d\Gamma^e$ differ for 3D, 2D and axisymmetric problems as given by (8.122)–(8.124), respectively. The tensor of hydraulic conductivity $\boldsymbol{K}^e$ of element e may be fully anisotropic in formulations introduced in Chap. 7. Is is important to note that the resulting discrete system of equations (9.20) is *symmetric* since the matrices $\boldsymbol{O}$ and $\boldsymbol{C}$ are symmetric.

Similarly, we obtain the matrices and RHS vector for the horizontal 2D flow in unconfined aquifers as

$$\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \int_{\Omega^e} (B^e S_o^e + \varepsilon_e^e)\, N_i N_j \, d\Omega^e \\
\boldsymbol{C} = C_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot (B^e \boldsymbol{K}^e \cdot \nabla N_j) d\Omega^e + \int_{\Gamma_C^e} B^e \Phi_h^e N_i N_j \, d\Gamma^e \Big) \\
\boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i \bar{Q}_h^e d\Omega^e + \int_{\Gamma_C^e} N_i B^e \Phi_h^e h_C^e \, d\Gamma^e - \\
&\qquad \int_{\Gamma_N^e} N_i B^e q_h^e d\Gamma^e \Big) - Q_w(t)\big|_i
\end{aligned} \tag{9.23}$$

where $B^e = h^e - f^{B^e}$ with $h^e = \sum_J N_J^e h_J^e$ and for the horizontal 2D flow in confined aquifers as

$$\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \int_{\Omega^e} \bar{S}_o^e \, N_i N_j \, d\Omega^e \\
\boldsymbol{C} = C_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{T}^e \cdot \nabla N_j) d\Omega^e + \int_{\Gamma_C^e} \bar{\Phi}_h^e N_i N_j \, d\Gamma^e \Big) \\
\boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i \bar{Q}_h^e d\Omega^e + \int_{\Gamma_C^e} N_i \bar{\Phi}_h^e h_C^e \, d\Gamma^e - \\
&\qquad \int_{\Gamma_N^e} N_i \bar{q}_h^e d\Gamma^e \Big) - Q_w(t)\big|_i
\end{aligned} \tag{9.24}$$

It is important to note the difference between (9.23) and (9.24) with respect to the used parameters. While for unconfined aquifer conditions (9.23) the input parameters S_o^e, $\boldsymbol{K}^e$, Φ_h^e and q_h^e are explicitly multiplied by the variable aquifer thickness $B^e = B^e(h^e)$ for each element e (except, however, the integral supply term $\bar{Q}_h^e$), for confined aquifer conditions these parameters have to be input as integral (already thickness-incorporating) values as $\bar{S}_o^e$, $\boldsymbol{T}^e$, $\bar{\Phi}_h^e$ and $\bar{q}_h^e$ of element e. An exception exists for unconfined aquifer conditions (9.23) in using integral Neumann-type and/or Cauchy-type BC's on Γ_N and Γ_D, respectively (Sect. 6.5.4). In this case the surface integrals of (9.23) become the same as in (9.24).

9.4 Time Integration

The matrix system (9.20) has to be solved in time t with the associated IC's via suitable single-step semi-implicit or fully implicit time marching recurrence schemes as described in Sect. 8.13. Most important are the GLS predictor-corrector time stepping method combined with an automatic error-controlled time step selection strategy (cf. Sect. 8.13.5 and Table 8.7)

$$\Big(\frac{\boldsymbol{O}}{\theta\Delta t_n}+\boldsymbol{C}\Big)\cdot\boldsymbol{h}_{n+1}=\boldsymbol{O}\cdot\Big[\frac{\boldsymbol{h}_n}{\theta\Delta t_n}+\big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{h}}_n\Big]+\boldsymbol{F}_{n+1} \tag{9.25}$$

where $\theta\in(\frac{1}{2},1)$ for the TR and BE scheme, respectively. On the other hand, for user-defined (fixed) time step sizes Δt_n the $\theta-$method (Sect. 8.13.4) is common

$$\Big(\frac{\boldsymbol{O}}{\Delta t_n}+\boldsymbol{C}\theta\Big)\cdot\boldsymbol{h}_{n+1}=\Big(\frac{\boldsymbol{O}}{\Delta t_n}-\boldsymbol{C}(1-\theta)\Big)\cdot\boldsymbol{h}_n+\big(\boldsymbol{F}_{n+1}\theta+\boldsymbol{F}_n(1-\theta)\big) \tag{9.26}$$

where $\theta\in(\frac{1}{2},\frac{2}{3},1)$ for the Crank-Nicolson, the Galerkin-in-time and the fully implicit scheme, respectively.

9.5 Free Surface Computation

9.5.1 Requirements

The treatment of free surfaces differs significantly between 2D and 3D problems. While the formulations of the basic equations for essentially horizontal problems in an unconfined (or phreatic) aquifer are rather simple, 3D problems can imply complicate conditions, especially due to

- The existence of multiple (more than one) free surfaces,
- Effects of location of free surface(s) on transport processes in the depth,
- Problems arising if parts of the domain of an aquifer system fall dry, and
- BC's and IC's become in dependence on the location of free surface(s).

We can formally distinguish between fixed-mesh and movable-mesh free-surface fully saturated modeling strategies (Fig. 9.1). The former strategy represents the classic method often previously preferred in groundwater modeling, which involves the inclusion of the entire flow domain in the analysis to achieve an allegedly robust computation. However, it requires a specific treatment of elements which are not or partially saturated. Unfortunately, such an approach can run into difficulties or can even fail. The main drawbacks concern:

1. The free-surface problem is commonly solved only in a non-rigorous manner, i.e., the kinematic BC (9.4) are adapted by ad-hoc approaches (e.g., by introducing an auxiliary 'well-term') such as done in the widely used finite-difference simulator MODFLOW [363]. Criticisms were summarized by Yeh et al. [581] and Knupp [313]. While Yeh et al. [581] modeled homogeneous 3D domains for which a moving technique is much simpler, Knupp [313] developed an improved

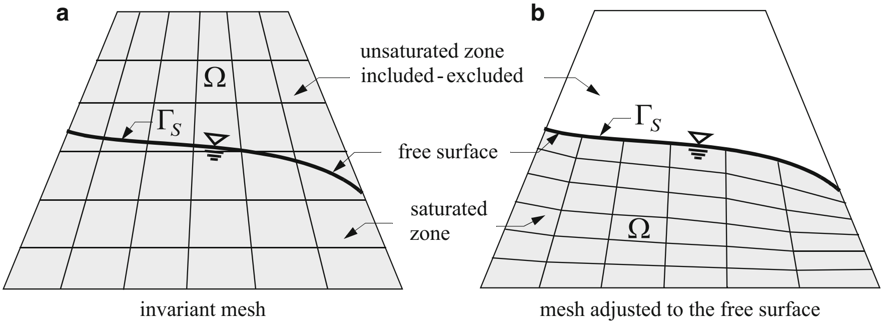


Fig. 9.1 (**a**) Fixed-mesh free-surface modeling strategy using an invariant (immovable) mesh and (**b**) fully saturated water-table modeling approach with moving (variable) mesh

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 fall dry. There are different 'tricks' to overcome such situations (e.g., frozen cells, converting procedures, intermediate deletion of elements). 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 associated with mass and heat transport processes, including density effects, forces to a generalization of the solution strategy.

Alternatively to a fixed-mesh strategy, the fully saturated modeling approach with 3D moving meshes has shown powerful and appropriate, provided a single coherent free surface in top position of an aquifer system exists. This approach considers only the domain below the free surface where the water table is treated as a moving material interface. It allows an accurate and rigorous modeling of both flow and transport processes.

9.5.2 Horizontal 2D Flow in Unconfined Aquifers

The GFEM approach to the vertically averaged equation (9.8) leads to a nonlinear matrix system (9.20) in the form

$$\boldsymbol{O}(\boldsymbol{h}) \cdot \dot{\boldsymbol{h}} + \boldsymbol{C}(\boldsymbol{h}) \cdot \boldsymbol{h} - \boldsymbol{F}(\boldsymbol{h}) = \boldsymbol{0} \tag{9.27}$$

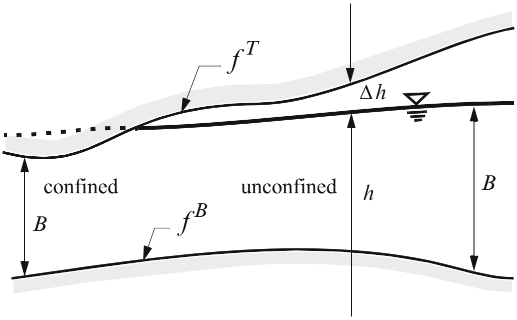


Fig. 9.2 Unconfined and confined conditions in an aquifer (*vertical cross section*)

and after applying time integration (9.25) or (9.26) in the form

$$\boldsymbol{A}(\boldsymbol{h}_{n+1}) \cdot \boldsymbol{h}_{n+1} = \boldsymbol{b}(\boldsymbol{h}_{n+1}, \boldsymbol{h}_n) \tag{9.28}$$

with

$$\boldsymbol{A}(\boldsymbol{h}_{n+1}) = \begin{cases} \dfrac{\boldsymbol{O}(\boldsymbol{h}_{n+1})}{\theta \Delta t_n} + \boldsymbol{C}(\boldsymbol{h}_{n+1}) & \text{predictor-corrector} \\ \dfrac{\boldsymbol{O}(\boldsymbol{h}_{n+1})}{\Delta t_n} + \boldsymbol{C}(\boldsymbol{h}_{n+1})\theta & \theta - \text{method} \end{cases} \tag{9.29}$$

and

$$\boldsymbol{b}(\boldsymbol{h}_{n+1}, \boldsymbol{h}_n) = \begin{cases} \boldsymbol{O}(\boldsymbol{h}_{n+1}) \cdot \big[\dfrac{\boldsymbol{h}_n}{\theta \Delta t_n} + \big(\tfrac{1}{\theta} - 1\big)\dot{\boldsymbol{h}}_n\big] + \boldsymbol{F}_{n+1}(\boldsymbol{h}_{n+1}) \\ \qquad\qquad\qquad\qquad \text{predictor-corrector} \\ \big(\dfrac{\boldsymbol{O}(\boldsymbol{h}_{n+1})}{\Delta t_n} - \boldsymbol{C}(\boldsymbol{h}_{n+1})(1-\theta)\big) \cdot \boldsymbol{h}_n + \\ \quad \boldsymbol{F}_{n+1}(\boldsymbol{h}_{n+1})\theta + \boldsymbol{F}_n(1-\theta) \qquad \theta - \text{method} \end{cases} \tag{9.30}$$

due to the $h-$dependency of the aquifer thickness $B = B(h) = h - f^B$ as recognized from (9.23), where the nonlinearities are shown in parentheses. The actually discharging aquifer thickness B differs for unconfined and confined conditions, viz.,

$$B = \begin{cases} h - f^B & \text{unconfined condition} \\ f^T - f^B & \text{confined confined} \end{cases} \tag{9.31}$$

where f^T and f^B are the top and bottom bounding surfaces, respectively, of the aquifer as sketched in Fig. 9.2.

The transition from the unconfined (phreatic) to the confined aquifer conditions is automatically realized in dependence on the computed hydraulic head h related to the top and bottom geometry of the aquifer, f^T and f^B, respectively. However, to prevent oscillatory effects at the transition state between confined and unconfined conditions, the computation of the effective storage coefficient $S = BS_o + \varepsilon_e$ is performed by using the smoothing *Heaviside* function

$$\begin{aligned} S &= BS_o + \varepsilon_e(1 - \varsigma) \\ \varsigma &= \tfrac{1}{\pi}\arctan(\tfrac{-\varDelta h}{\sigma}) + \tfrac{1}{2} \end{aligned} \tag{9.32}$$

where $\varDelta h = f^T - h$ (Fig. 9.2) is the difference between the aquifer top and the computed head h, and $\sigma \approx 10^{-3}$ represents a given smoothing parameter. Accordingly, for $\varDelta h \geq 0$ unconfined and for $\varDelta h < 0$ confined conditions occur.

The iterative solution (cf. Sect. 8.18) of the nonlinear system (9.28) is performed either by a common Picard iteration method of linear convergence rate

$$\boldsymbol{A}(\boldsymbol{h}_{n+1}^{\tau}) \cdot \boldsymbol{h}_{n+1}^{\tau+1} = \boldsymbol{b}(\boldsymbol{h}_{n+1}^{\tau}, \boldsymbol{h}_n) \quad \tau = 0, 1, 2, \ldots \tag{9.33}$$

or via a full Newton iteration method of quadratic convergence rate

$$\big[\boldsymbol{A}(\boldsymbol{h}_{n+1}^{\tau}) + \hat{\boldsymbol{J}}(\boldsymbol{h}_{n+1}^{\tau})\big] \cdot \boldsymbol{h}_{n+1}^{\tau+1} = \hat{\boldsymbol{J}}(\boldsymbol{h}_{n+1}^{\tau}) \cdot \boldsymbol{h}_{n+1}^{\tau} + \boldsymbol{b}(\boldsymbol{h}_{n+1}^{\tau}, \boldsymbol{h}_n) \quad \tau = 0, 1, 2, \ldots \tag{9.34}$$

with the partial Jacobian

$$\begin{aligned} \hat{\boldsymbol{J}}(\boldsymbol{h}_{n+1}^{\tau}) &= \frac{\partial \boldsymbol{A}(\boldsymbol{h}_{n+1}^{\tau})}{\partial \boldsymbol{h}_{n+1}^{\tau}} \cdot \boldsymbol{h}_{n+1}^{\tau} - \frac{\partial \boldsymbol{b}(\boldsymbol{h}_{n+1}^{\tau}, \boldsymbol{h}_n)}{\partial \boldsymbol{h}_{n+1}^{\tau}} \\ \hat{J}_{ij} &= \sum_l \frac{\partial A_{il}}{\partial h_j^{\tau}} h_l^{\tau} - \frac{\partial b_i}{\partial h_j^{\tau}} \end{aligned} \tag{9.35}$$

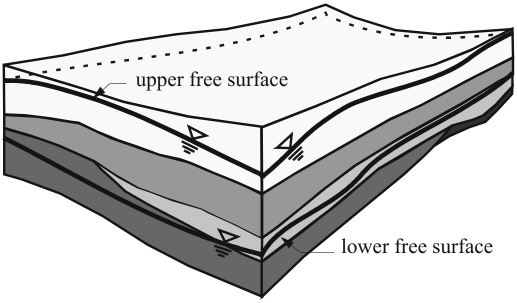
where τ is an iteration counter and the partial Jacobian is given by

$$\hat{J}_{ij} = \begin{cases} \frac{1}{\theta \varDelta t_n} \hat{J}_{ij}^a + \hat{J}_{ij}^b - \frac{1}{\theta \varDelta t_n} \hat{J}_{ij}^c - (\frac{1}{\theta} - 1) \hat{J}_{ij}^d - \hat{J}_{ij}^f & \text{predictor-corrector} \\ \frac{1}{\varDelta t_n} \hat{J}_{ij}^a + \theta \hat{J}_{ij}^b - \frac{1}{\varDelta t_n} \hat{J}_{ij}^c + (1 - \theta) \hat{J}_{ij}^e - \theta \hat{J}_{ij}^f & \theta - \text{method} \end{cases} \tag{9.36}$$

with

$$\begin{aligned} \hat{J}_{ij}^a &= \textstyle\sum_e \int_{\varOmega^e} S_o^e \, N_i N_j \sum_l (N_l h_{l,n+1}^{\tau}) d\varOmega^e \\ \hat{J}_{ij}^b &= \textstyle\sum_e \Big(\int_{\varOmega^e} \nabla N_i \cdot \big(\boldsymbol{K} N_j \cdot \sum_l (\nabla N_l h_{l,n+1}^{\tau})\big) d\varOmega^e + \int_{\varGamma_C^e} \varPhi_h^e N_i N_j \sum_l (N_l h_{l,n+1}^{\tau}) d\varGamma^e \Big) \\ \hat{J}_{ij}^c &= \textstyle\sum_e \int_{\varOmega^e} S_o^e \, N_i N_j \sum_l (N_l h_{l,n}) d\varOmega^e \\ \hat{J}_{ij}^d &= \textstyle\sum_e \int_{\varOmega^e} S_o^e \, N_i N_j \sum_l (N_l \dot{h}_{l,n}) d\varOmega^e \\ \hat{J}_{ij}^e &= \textstyle\sum_e \Big(\int_{\varOmega^e} \nabla N_i \cdot \big(\boldsymbol{K} N_j \cdot \sum_l (\nabla N_l h_{l,n})\big) d\varOmega^e + \int_{\varGamma_C^e} \varPhi_h^e N_i N_j \sum_l (N_l h_{l,n}) d\varGamma^e \Big) \\ \hat{J}_{ij}^f &= \textstyle\sum_e \Big(\int_{\varGamma_C^e} N_i N_j \varPhi_h^e h_C^e d\varGamma^e - \int_{\varGamma_N^e} N_i N_j q_h^e d\varGamma^e \Big) \end{aligned} \tag{9.37}$$

Fig. 9.3 Free-surface flow in a 3D aquifer system



We note that the Picard method preserves symmetry of the discrete system of equations, while the Newton method generates an unsymmetric system matrix due to the unsymmetry of the partial Jacobian $\hat{\boldsymbol{J}}$, which increases the computational effort for the Newton method. In using the GLS predictor-corrector time integration with automatic time step control a one-step Picard or one-step Newton method is usually preferred (see Sect. 8.18.4), in which no iteration per time step is executed and the iterate $\boldsymbol{h}_{n+1}^{\tau}$ appearing in (9.33)–(9.37) is replaced by the predictor solution $\boldsymbol{h}_{n+1}^{p}$. On the other hand, for steady-state solutions ($\varDelta t_n \to \infty$) and for transient solutions with fixed (predefined) time step sizes an iterative cycling is always required until convergence is achieved

$$\frac{\|\boldsymbol{h}_{n+1}^{\tau+1} - \boldsymbol{h}_{n+1}^{\tau}\|}{\|\boldsymbol{h}_{n+1}^{\tau+1}\|} \leq \epsilon \tag{9.38}$$

where ϵ is the error tolerance to be defined.

9.5.3 3D Free-Surface Flow Modeling with Moving Meshes and BASD Technique

In a 3D aquifer system (Fig. 9.3) involving a coherent free surface on top (and possibly further free surfaces in deeper locations) the governing flow equation (9.2) becomes nonlinear due to the kinematic free-surface BC (9.4) consisting of two parts which have to be satisfied simultaneously on $\varGamma_S$: (1) a Neumann-type flux condition $-(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} = \varepsilon_e \frac{\partial h}{\partial t} - P|_{\varGamma_S}$ and (2) a constant pressure condition occurring as a Dirichlet-type condition $h = z|_{\varGamma_S}$. While the first condition is already built in the $\varGamma_S$−surface integral appearing in the finite element formulation (9.20) with (9.22), it is necessary to find suitable solution strategies for adapting the vertical location of the free surface to the head solution to satisfy $z = h|_{\varGamma_S}$. An appropriate iterative method is based on vertically moving meshes.

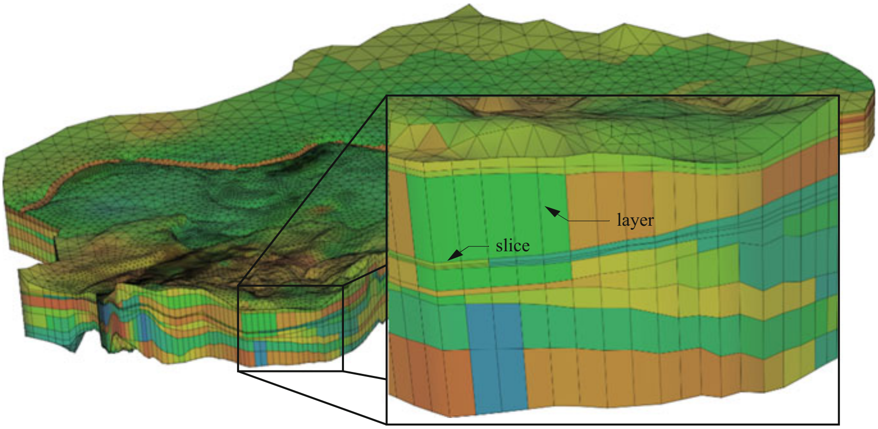


Fig. 9.4 Prismatic element mesh of a layered aquifer structure

Suppose that a 3D finite element mesh is aligned to a stratigraphic structure formed by a number of layers and slices and discretized by prismatic elements (Fig. 9.4, Sect. 15.1.4). The adaptation of such a mesh to a changeable *a priori* unknown free-surface location has the advantage that the mesh density is maintained in domains which are actually discharged. For vertically moving meshes in a fully saturated modeling approach an accurate and powerful technique is required to adjust not only the top slice to the free surface location, also the slices of the inner mesh have to be suitably adapted in accordance with the changed free surface location to ensure a vertical well-spaced mesh density during mesh shrinkage or expansion and to avoid any layer intersection. As a consequence of the vertical mesh movement in a stratigraphical aquifer structure a number of data must be adapted to the new locations of slices. Material parameters (e.g., conductivities and storage coefficients) are here of specific concern because they can abruptly change from layer to layer and possess high contrasts. It is obvious that interpolation of those parameters onto the new mesh coordinates can smooth parameter discontinuities at stratigraphic interfaces. Accordingly, a technique is needed which effectively adapts and carefully reassigns the material parameters at a minimum of parameter interpolation.

This technique termed BASD (*Best-Adaptation-to-Stratigraphic-Data*) transforms and joins the model data containing the stratigraphic initial structure to a moving finite element mesh which is appropriately adapted to the free-surface locations. In this adaptation process the mesh slices are aligned in such a manner that the adjusted mesh is exactly fitted to parameter discontinuities if they ever exist. Remaining slices can be shifted and repositioned to achieve a well-spaced nodal distribution in the depth without unnecessary mesh refinement and coarsening if ever attainable. Following criteria are employed to position the material layers:

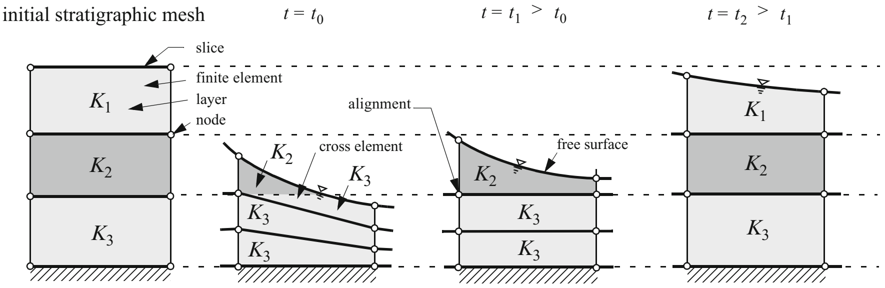


Fig. 9.5 Moving mesh BASD technique of parameter adaptation applied to a 3D free-surface problem: schematized example for a groundwater table rise in time t

- Fit layer interfaces (slices) to interfaces of parameter discontinuity once located within the domain,
- Distribute remaining slices proportional to layer thicknesses to attain as best as possible well-behaved mesh distances in the depth, and
- Join parameter for such elements according to partial volumes and surfaces, which cross and intersect more than one stratigraphic material layer.

The principle of the mesh adaptation process is sketched in Fig. 9.5. The initial stratigraphy consists of three layers with different isotropic conductivities K_1, K_2 and K_3. At the initial time t_0 the water table h 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 required here. Such type of 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. 9.5 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:

1. Compute the hydraulic head $\boldsymbol{h}_{n+1}$ at the new time plane by solving (9.25) or (9.26) with (9.22).
2. Determine a new free surface location for the upper slice $s = \text{top} = 1$ of the moving mesh

$$\boldsymbol{z}_{n+1}^{\text{top}} = \boldsymbol{h}_{n+1}(\boldsymbol{x}, t) \tag{9.39}$$

satisfying the head condition of (9.4), where $\boldsymbol{z}_{n+1}^{\text{top}}$ corresponds to z−coordinates of the top slice.

3. Adjust and distribute the inner slices, indexed by s, according to

$$\begin{aligned} z_{n+1}^{s} &= z^{\text{rig}} + L^{\text{rel}}(z_n^s - z^{\text{rig}}), \quad s = 2, 3, \ldots, \text{rig} - 1; \quad s = 1 \Rightarrow \text{top} \\ L^{\text{rel}} &= (z_{n+1}^{\text{top}} - z^{\text{rig}})/(z_n^{\text{top}} - z^{\text{rig}}) \end{aligned} \tag{9.40}$$

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

$$z_{n+1}^{s} = z_{n+1}^{s+1} + \tfrac{1}{n_d + n_h}(z_{\text{upper}}^{\text{rig}} - z_{\text{lower}}^{\text{rig}}) \tag{9.41}$$

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

4. Assign the parameter arrays according to the new layer positions. Two cases are distinguished: (a) achievement of full alignment (no interpolation) and (b) data interpolation and joining for so-called *cross elements* (Fig. 9.5).
5. Find out cross elements and join their properties. The joining process differs between volume-specified materials (such as conductivity $\boldsymbol{K}$, storage coefficient S_o) and area-specified data (such as specific yield ε_e). For volume-specified material data Gauss-Legendre numerical integration (cf. Sect. 8.12) is used to determine the partial volumes Ω_i^e of a finite element e intersecting the stratigraphic contours. The material property K^e of such a cross element is computed by a partial volume-weighted average as

$$K^e = \frac{1}{\Omega^e} \sum_{i=1}^{N} \Omega_i^e K_i^e \tag{9.42}$$

where N is the number of intersections and K_i^e are the properties of the intersected layer.

Similarly, a partial area-weighted averaging process is preferred for areal properties, e.g., specific yield ε_e^e or infiltration rate P^e of element e. However, it has been found a numerical integration is here insufficient. Analytical formulae have been developed to determine exactly the intersected areas of an element. It leads to a telescoping sum to average an areal property A^e in the form

$$\begin{aligned} A^e = \sum_{i=1}^{l} (\lambda_{12}^{i}\lambda_{13}^{i} - \lambda_{12}^{i-1}\lambda_{13}^{i-1})A_i^e + [\lambda_{31}^{l}(1 - \lambda_{32}^{l})]A_l^e + \\ \sum_{i=l+1}^{N+1} (\lambda_{31}^{i-1}\lambda_{32}^{i-1} - \lambda_{31}^{i}\lambda_{32}^{i})A_i^e \end{aligned} \tag{9.43}$$

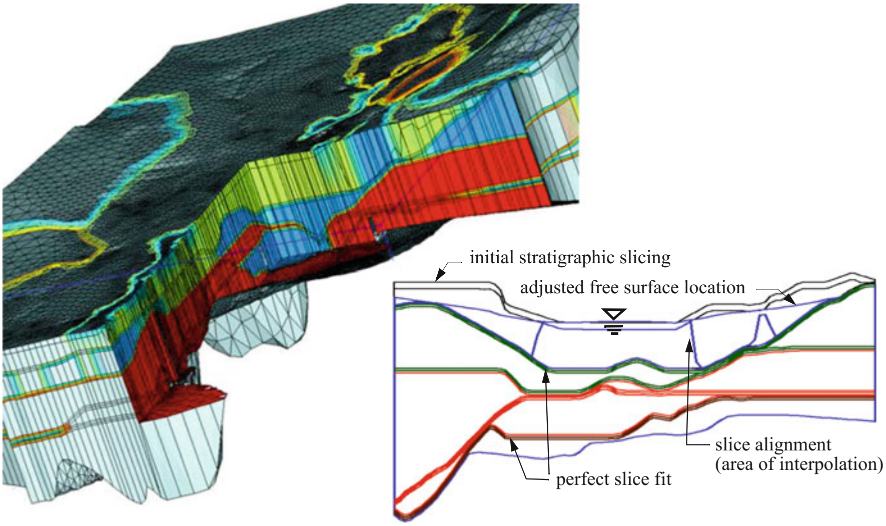


Fig. 9.6 Moving mesh BASD technique applied to a complex 3D stratigraphy: 3D model cut view (distribution of conductivities) and moving mesh along a cross section (initial stratigraphy versus adapted slicing)

with the weights

$$\lambda_{mn}^{i} = \frac{1}{1 - \dfrac{h_n - z_n^i}{h_m - z_m^i}} \qquad i = 1, \ldots, N; \quad m = 1, 3; \quad n = 1, 2, 3 \tag{9.44}$$

and the definitions

$$\lambda_{mn}^{0} = 0 \quad \text{and} \quad \lambda_{mn}^{N+1} = 1 \tag{9.45}$$

written for triangular top and bottom areas of a prismatic pentahedral element, where A^e and A_i^e are the averaged and partial areal properties, respectively, h_n and z_n^i correspond to the hydraulic head and the ith stratigraphic z–coordinates at (local) node n, respectively, and the index l represents the lth intersected layer for which the partial area is not a triangle, generally a pentagonal area. Equivalent averages are used for hexahedral elements, where each of their quadrilateral top and bottom areas are subdivided into four triangles.

The use of the BASD technique for a complex multi-aquifer system is illustrated in Fig. 9.6. It reveals how the mesh fits and moves through the complex stratigraphy consisting of a number of aquifers and aquitards. To satisfy the condition (9.39) in a moving mesh the final matrix system (9.20) with (9.22), (9.25) and (9.26) becomes nonlinear

$$\boldsymbol{A}(\boldsymbol{h}_{n+1}) \cdot \boldsymbol{h}_{n+1} = \boldsymbol{b}(\boldsymbol{h}_{n+1}, \boldsymbol{h}_n) \tag{9.46}$$

for solving $\boldsymbol{h}_{n+1}$ at the new time plane $n+1$, where

$$\boldsymbol{A}(\boldsymbol{h}_{n+1}) = \begin{cases} \dfrac{\boldsymbol{O}(\boldsymbol{h}_{n+1})}{\theta \Delta t_n} + \boldsymbol{C}(\boldsymbol{h}_{n+1}) & \text{predictor-corrector} \\ \dfrac{\boldsymbol{O}(\boldsymbol{h}_{n+1})}{\Delta t_n} + \boldsymbol{C}(\boldsymbol{h}_{n+1})\theta & \theta - \text{method} \end{cases} \tag{9.47}$$

and

$$\boldsymbol{b}(\boldsymbol{h}_{n+1}, \boldsymbol{h}_n) = \begin{cases} \boldsymbol{O}(\boldsymbol{h}_{n+1}) \cdot \big[\dfrac{\boldsymbol{h}_n}{\theta \Delta t_n} + \big(\frac{1}{\theta} - 1\big)\dot{\boldsymbol{h}}_n\big] + \boldsymbol{F}_{n+1}(\boldsymbol{h}_{n+1}) \\ \qquad\qquad\qquad\qquad\qquad \text{predictor-corrector} \\ \big(\dfrac{\boldsymbol{O}(\boldsymbol{h}_{n+1})}{\Delta t_n} - \boldsymbol{C}(\boldsymbol{h}_{n+1})(1-\theta)\big) \cdot \boldsymbol{h}_n + \\ \quad \boldsymbol{F}_{n+1}(\boldsymbol{h}_{n+1})\theta + \boldsymbol{F}_n(1-\theta) \qquad \theta - \text{method} \end{cases} \tag{9.48}$$

The solution of (9.46) is commonly performed via a Picard iteration method. In the case of using predefined time step sizes Δt_n in transient flow or solving steady-state flow problems, the iteration τ occurs

$$\boldsymbol{A}(\boldsymbol{h}_{n+1}^{\tau}) \cdot \boldsymbol{h}_{n+1}^{\tau+1} = \boldsymbol{b}(\boldsymbol{h}_{n+1}^{\tau}, \boldsymbol{h}_n) \quad \tau = 0, 1, 2, \ldots \tag{9.49}$$

until convergence is achieved

$$\frac{\|\boldsymbol{h}_{n+1}^{\tau+1} - \boldsymbol{h}_{n+1}^{\tau}\|}{\|\boldsymbol{h}_{n+1}^{\tau+1}\|} \le \epsilon \tag{9.50}$$

where ϵ is a defined error tolerance. On the other hand, in using the predictor-corrector time integration with automatic error-controlled time stepping a one-step Picard method for transient free-surface flow problems is preferred, in which the predictor solution $\boldsymbol{h}_{n+1}^{p}$ is used to linearize (9.49) in the form

$$\boldsymbol{A}(\boldsymbol{h}_{n+1}^{p}) \cdot \boldsymbol{h}_{n+1} = \boldsymbol{b}(\boldsymbol{h}_{n+1}^{p}, \boldsymbol{h}_n) \tag{9.51}$$

Due to the Picard iteration method the resulting discrete systems in form of (9.46) or (9.49) remain symmetric.

9.5.4 3D Free-Surface Flow Modeling with Fixed Meshes and Pseudo-unsaturated Conditions

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 flow modeling in unconfined aquifers to use exclusively fixed grids (e.g., [170,363]). Fixed grid techniques have to mimic, more or less, unsaturated flow conditions to control the solution process for saturated, partly saturated or completely dry mesh elements. Since a physically true unsaturated flow approach is avoided, such kind of unsaturated flow modeling, here termed as *pseudo-unsaturated* flow, represents only a physical approximation and quite different forms of implementation can be found in the literature (see discussion in [313]). 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, in [27] the conductivity K is assigned to a very small constant value as soon the pressure head ψ (3.259) becomes negative: $K/1{,}000$ for $\psi < 0$ and K if $\psi \geq 0$. Apparently, this is a crude controlling procedure since it does not differ between the degrees of saturation of the elements. Desai and Li [123] have improved the technique for finite elements by introducing linear relationships of conductivity and storage coefficient 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 [123], 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^e is determined from the actual filling height ψ of fluid in an element e (Fig. 9.7b):

$$s_p^e = s_p^e(\psi) = \frac{\Omega^{ef}(\psi)}{\Omega^e} \tag{9.52}$$

and the integration of element balance terms is only accomplished over the fluid-filled volume Ω^{ef} of element e, viz.,

$$\int_{\Omega^{ef}} (.) d\Omega^e \approx \int_{\Omega^e} (.) s_p^e \, d\Omega^e \tag{9.53}$$

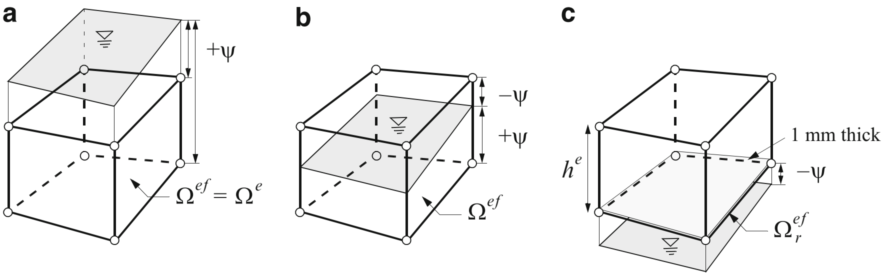


Fig. 9.7 Three cases of pseudo-saturation s_p^e: (**a**) saturated, (**b**) partially saturated and (**c**) fully unsaturated (dry) element *e*

Accordingly, the pseudo-saturation s_p^e becomes related to the actual geometric condition of the used spatial discretization. It provides a geometry-consistent scaling of balance terms and has proved superior to a simple parameter-switching as stated above. Three cases can be distinguished (Fig. 9.7):

- An element e is considered saturated if $\psi \geq 0$ at all nodes of the element. Then it becomes $\Omega^{ef} = \Omega^e$ and $s_p^e \equiv 1$.
- 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)).
- An element e is considered fully unsaturated (or dry) if $\psi < 0$ at all related element nodes. Since Ω^{ef} have to be positive and non-zero the volume must be constrained by a minimum. Practically, a minimum filling height (e.g., 1 mm) is employed to limit Ω^{ef} by Ω_r^{ef}. This leads to a measure of a residual pseudo-saturation $s_{pr}^e = \Omega_r^{ef}/\Omega^e > 0$ for such an element:

$$0 < s_{pr}^e \leq s_p^e \leq 1, \quad 0 < \Omega_r^{ef} \leq \Omega^{ef} \leq \Omega^e \tag{9.54}$$

Hence, the pseudo-saturation s_p^e for each element represents a linear relationship of the pressure head ψ (Fig. 9.8), viz.,

$$s_p^e = \begin{cases} 1 + \dfrac{(1 - s_{pr}^e)\psi}{h^e} & \text{for} \quad -h^e < \psi < 0 \\ 1 & \text{for} \quad \psi \geq 0 \\ s_{pr}^e & \text{for} \quad \psi \leq -h^e \end{cases} \tag{9.55}$$

where h^e is the height of element e. The expression (9.55) is similar to the linear relationship (D.19) of Appendix D used for true unsaturated problem formulations. The difference of (9.55), however, is in using a capillary fringe thickness of element height h^e, which makes the pseudo-saturation dependent on the spatial discretization.

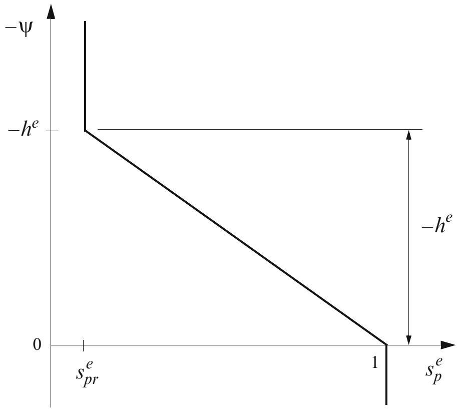


Fig. 9.8 Linear relationship between pressure head ψ and pseudo-saturation s^e_p of element e with height h^e

Using (9.52) and (9.53) in the finite element equations (9.22) it leads to a natural approach for evaluating the corresponding integral terms in a weak solution:

$$\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \Big(\int_{\Omega^e} S^e_o \, N_i N_j \, s^e_p \, d\Omega^e + \int_{\Gamma^{ef}_S} \varepsilon^e_e \, N_i N_j \, d\Gamma^e \Big) \\
\boldsymbol{C} = C_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{K}^e \cdot \nabla N_j) \, s^e_p \, d\Omega^e + \int_{\Gamma^{ef}_C} \Phi^e_h N_i N_j \, d\Gamma^e \Big) \\
\boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i \, Q^e_h \, s^e_p \, d\Omega^e + \int_{\Gamma^{ef}_C} N_i \Phi^e_h h^e_C \, d\Gamma^e - \int_{\Gamma^{ef}_N} N_i q^e_h d\Gamma^e + \\
&\qquad \int_{\Gamma^{ef}_S} N_i \, P^e \, d\Gamma^e \Big) - Q_w(t)\Big|_i
\end{aligned} \tag{9.56}$$

By using (9.56) the solution of the resulting matrix system is equivalent to (9.46)–(9.51) as described in the previous Sect. 9.5.3. Note that the surface integrals over Γ^{ef}_S, Γ^{ef}_N and Γ^{ef}_C appearing in (9.56) have to be evaluated in accordance with the actual filling height ψ occurring in the corresponding element. The free surface integral over Γ^{ef}_S only exists in partially saturated elements where the phreatic surface lies in the interior of an element volume (this integral has to be evaluated for a surface which is spanned by the $\psi-$heights), otherwise it is dropped for all saturated and dry elements. On the other hand, BC-related surface integrals Γ^{ef}_N and Γ^{ef}_C are generally not applied to dry elements.

It should be emphasized that 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 usually inferior to a moving mesh strategy with respect to the attainable accuracy.

9.6 Incorporation of Multilayer Well Flow BC

For 3D flow problems there is the need for modeling well-type SPC's (9.3) at nodal points of well discharge in form of multilayer well BC's as introduced in Sect. 6.5.6. In the context of FEM multilayer well BC's can be easily modeled as 1D tubular discrete features. Then, the governing flow equation of pure homogeneous liquid in a well bore is, cf. (6.77) and Table 4.5,

$$A_w S_{ow} \frac{\partial h}{\partial t} - A_w \nabla \cdot (K_w \nabla h) = -Q_w \delta(s - s_w)\,, \quad w = 1, \ldots, N_W \tag{9.57}$$

with the specific storage coefficient of the well

$$S_{ow} = \begin{cases} \frac{1}{L_w} + \rho_0 g \gamma & \text{phreatic} \\ \rho_0 g \gamma & \text{non-phreatic} \end{cases} \tag{9.58}$$

the cross-sectional flow area of the well bore

$$A_w = \pi R^2 \tag{9.59}$$

and

$$K_w = \frac{R^2 \rho_0 g}{8 \mu_0} \tag{9.60}$$

by using the Hagen-Poiseuille law (4.51), where $\nabla = \partial/\partial s$ is defined here for the 1D line direction s along the well axis, R is the well radius and L_w is the length of the liquid-filled well bore (other related variables are defined in Sect. 6.5.6 and Chap. 4). Applying the GFEM procedure to (9.57) similar to the derivation of the 1D ADE as done in Sect. H.1 of Appendix H, discretizing each well w by a number of 1D tubular finite elements called *discrete feature elements* (DFE's), cf. Chap. 14, as sketched in Fig. 9.9 and superimposing the contributions for sharing nodes connected to both the 3D porous-medium elements and the 1D well DFE's, the formulation of the finite element matrices and RHS vector of (9.22) extends now to

$$\begin{aligned} \boldsymbol{O} = O_{ij} &= \sum_e \Big(\int_{\Omega^e} S_o^e\, N_i N_j \, d\Omega^e + \int_{\Gamma_S^e} \varepsilon_e^e\, N_i N_j \, d\Gamma^e + \int_{S_w^e} A_w^e S_{ow}^e\, N_i N_j \, dS^e \Big) \\ \boldsymbol{C} = C_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{K}^e \cdot \nabla N_j) d\Omega^e + \int_{\Gamma_C^e} \Phi_h^e N_i N_j \, d\Gamma^e + \\ &\qquad \int_{S_w^e} A_w^e \nabla N_i \cdot (K_w^e \nabla N_j) dS^e \Big) \\ \boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i Q_h^e d\Omega^e + \int_{\Gamma_C^e} N_i \Phi_h^e h_C^e \, d\Gamma^e - \int_{\Gamma_N^e} N_i q_h^e d\Gamma^e + \\ &\qquad \int_{\Gamma_S^e} N_i P^e d\Gamma^e \Big) - Q_w(t)\big|_i \end{aligned} \tag{9.61}$$

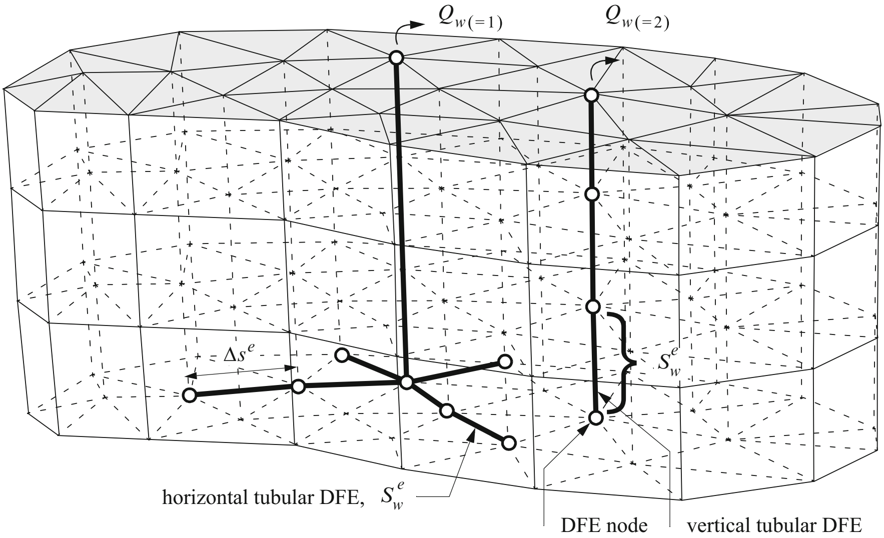


Fig. 9.9 Representation of a vertical and horizontal well configuration in a 3D discretization by using 1D tubular DFE

where the summation of e runs over all elements, including the 1D DFE's of all incorporated wells w, and S_w^e corresponds to a line tubular DFE segment of well w. We note that the basis functions N_i vary for 3D elements, Ω^e, and 1D DFE's, S_w^e.The additional matrix contributions in $\boldsymbol{O}$ and $\boldsymbol{C}$ of (9.61) for the 1D well-type DFE's can be analytically determined as demonstrated in Sect. H.1 of Appendix H. We obtain for each DFE segment e

$$\begin{aligned}\int_{S_w^e} A_w^e S_{ow}^e \, N_I N_J dS^e &= \frac{A_w^e S_{ow}^e \Delta s^e}{6}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\\ \int_{S_w^e} A_w^e \nabla N_I \cdot (K_w^e \nabla N_J) dS^e &= \frac{A_w^e K_w^e}{\Delta s^e}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\end{aligned} \qquad (9.62)$$

where Δs^e is the line segment length of the 1D DFE e. In practical applications the storage in the well casing can often be neglected since $A_w^e S_{ow}^e$ is usually small. The well conductivity K_w (9.60) implies a high parameter contrast, typically $K_w \approx \mathcal{O}(10^6)\ \mathrm{m\,s^{-1}}$, which ensures a relatively uniform hydraulic head h along the nodes forming the well axis producing only a slight gradient in the well tube toward the node representing the exit point of well discharge Q_w (at the pump position).

9.7 Computation of Darcy Velocities and Flow Budget Analysis

The computation of the Darcy velocity $\boldsymbol{q} = -\boldsymbol{K} \cdot \nabla h$ for 3D flow (9.1) and similarly with (9.6) and (9.11) for horizontal 2D flow problems is performed in the discrete form (cf. Sect. 8.19.1)

$$\boldsymbol{q}(\boldsymbol{x}, t_{n+1}) = -\sum_j \boldsymbol{K} \cdot \nabla N_j(\boldsymbol{x})\, h_j(t_{n+1}) \tag{9.63}$$

based on the known hydraulic head $h_j(t_{n+1}) = \boldsymbol{h}_{n+1}$ which has been solved at each nodal point $j = 1, \ldots, N_{\mathrm{P}}$ from (9.25) or (9.26) at the new time plane $n+1$. The evaluation of (9.63) is combined with appropriate smoothing techniques such as described in Sect. 8.19.1 to obtain Darcy velocities at the nodal points. A typical Darcy velocity field is exhibited in Fig. 9.10a computed by using superconvergent flux evaluation and local smoothing (see Sect. 8.19.1.2) providing a precise and continuous representation of the nodal flow vectors.

On the other hand, the flux computation at boundary sections for the purpose of a precise flow budget analysis is done via the CBFM as introduced and thoroughly described in Sect. 8.19.2. The corresponding weak formulation for solving the consistent boundary flux q_n at any exterior or interior boundary section $\varGamma$ is given for 3D flow as

$$\int_\varGamma N_i\, q_n\, d\varGamma = -\int_\varOmega N_i S_o \frac{\partial h}{\partial t} d\varOmega - \int_\varOmega \nabla N_i \cdot (\boldsymbol{K} \cdot \nabla h) d\varOmega + \int_\varOmega N_i (Q_h + Q_{hw}) d\varOmega \tag{9.64}$$

where $h = \sum_j N_j\, h_j$ is known at t_{n+1}. The consistent boundary flux vector $\boldsymbol{q}_n$ is now solved from the resulting matrix system

$$\boldsymbol{M} \cdot \boldsymbol{q}_n = -\boldsymbol{O}^\dagger \cdot \dot{\boldsymbol{h}} - \boldsymbol{C}^\dagger \cdot \boldsymbol{h} + \boldsymbol{F}^\dagger \tag{9.65}$$

where

$$\begin{aligned}
\boldsymbol{M} &= M_{ij} = \textstyle\int_\varGamma N_i N_j \, d\varGamma \\
\boldsymbol{O}^\dagger &= O_{ij}^\dagger = \textstyle\int_\varOmega S_o N_i N_j \, d\varOmega \\
\boldsymbol{C}^\dagger &= C_{ij}^\dagger = \textstyle\int_\varOmega \nabla N_i \cdot (\boldsymbol{K} \cdot \nabla N_j) d\varOmega \\
\boldsymbol{F}^\dagger &= F_i^\dagger = \textstyle\int_\varOmega N_i Q_h d\varOmega - Q_w(t)\big|_i
\end{aligned} \tag{9.66}$$

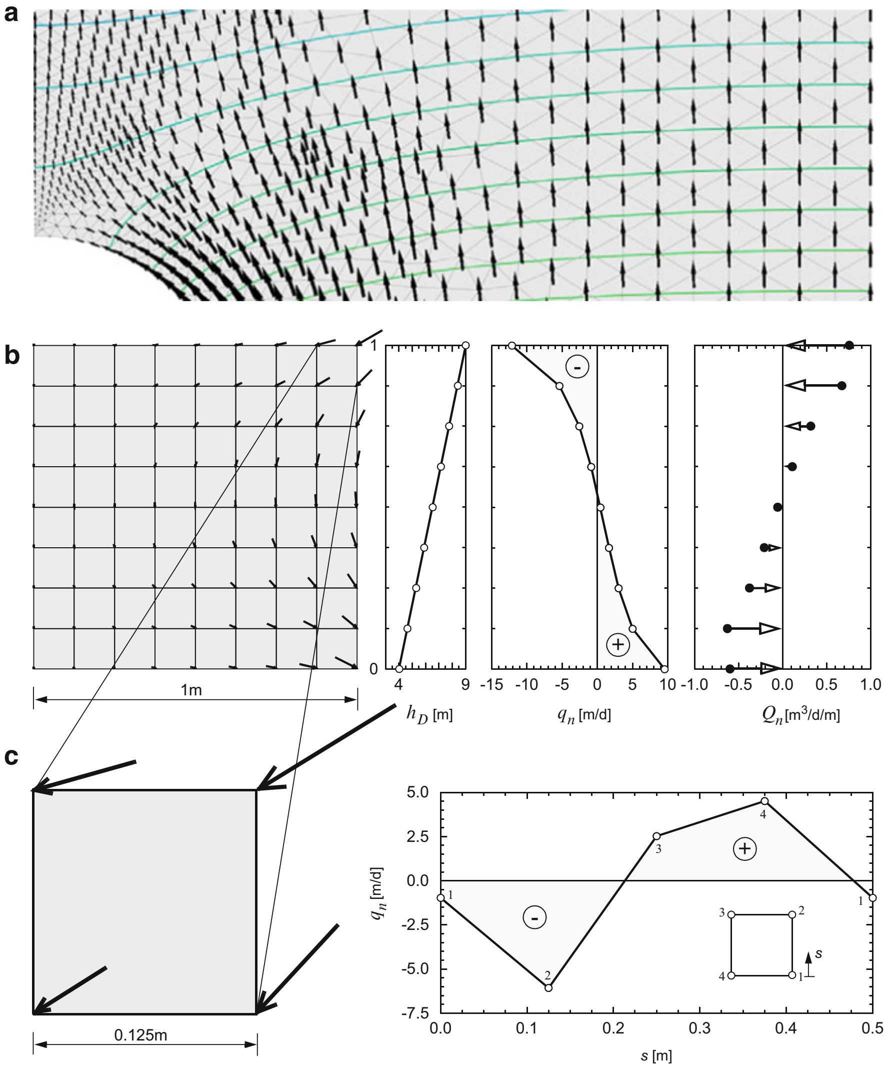


Fig. 9.10 (**a**) Darcy velocity $\boldsymbol{q}$ field with hydraulic head contours h for a corner flow situation (cut-out of triangle mesh), (**b**) balanced boundary flux q_n and integral balance flux Q_n evaluated at a Dirichlet-type boundary of given h_D on the right vertical side of a closed unit-square box steady-state flow problem (at $K = 1\ \mathrm{m\,d^{-1}}$ and without sources/sinks) with drawn $\boldsymbol{q}$ and (**c**) balanced boundary flux q_n on the enclosing boundary of a single finite element separated from the upper entry corner to indicate local conservativity $\int_{\Gamma^e} q_n d\Gamma = 0$

exemplified for 3D flow. Similar formulations result for the 2D horizontal flow equations. Alternatively, the integral boundary balance flux $\boldsymbol{Q}_n$ can be directly evaluated at each boundary node, viz.,

$$\begin{aligned} \boldsymbol{Q}_n &= -\boldsymbol{M} \cdot \boldsymbol{q}_n \\ &= \boldsymbol{O}^{\dagger} \cdot \dot{\boldsymbol{h}} + \boldsymbol{C}^{\dagger} \cdot \boldsymbol{h} - \boldsymbol{F}^{\dagger} \end{aligned} \tag{9.67}$$

where $\dot{\boldsymbol{h}}$ and $\boldsymbol{h}$ are known at the corresponding evaluation time t_{n+1}. Typical q_n and Q_n are illustrated in Fig. 9.10b,c for a simple unit-square box steady-state flow problem, where a linearly distributed Dirichlet BC is imposed on the right vertical side of the otherwise closed box. The computed q_n satisfies exactly mass conservation at the global boundary by $\int_{\Gamma} q_n = 0$, where the magnitude of mass entering the domain through the boundary Γ is in equilibrium with the magnitude of mass leaving the domain through the same boundary Γ. Due to the relatively coarse mesh, we observe that the nodal velocity plot $\boldsymbol{q}$ must not be locally consistent with the balance flux. This is particularly evident in the element at the upper entry corner of the box as magnified in Fig. 9.10c. While the approximation of $\boldsymbol{q}$ is limited by its elementwise constant behavior in linear elements, nevertheless, the boundary flux q_n along the enclosing boundary Γ^e of the single element satisfies local conservativity by $\int_{\Gamma^e} q_n = 0$.

9.8 Examples

9.8.1 Transient Flow to a Well in a Confined and Unconfined Aquifer

Les us consider a fully penetrating single pumping well which extracts water from a porous aquifer at a constant rate. As a result, the pumping well induces a transient lowering of the water table termed *drawdown* in the vicinity of the well causing a cone of depression as sketched in Fig. 9.11 for both confined and unconfined conditions. We assume flow regimes in the confined and unconfined aquifer, which provide comparable discharging thicknesses.

For the case of a *confined* aquifer (Fig. 9.11a) there is an exact analytical solution firstly presented by Theis (1935) [34, 511]. Theis' solution is associated with the fundamental assumptions that the confined aquifer has an infinite areal extent and a uniform thickness of homogeneous and isotropic porous material. Furthermore, Dupuit assumption (Sect. 3.5) must be valid, the confined aquifer has a constant thickness and the aquifer bottom is horizontal, the well discharge is constant, the well penetrates the entire thickness and well storage effects are negligible. Under these conditions the Theis' analytical solution is [34]

$$h(r,t) = h_0 - \frac{Q_w}{4\pi T} W(u) \tag{9.68}$$

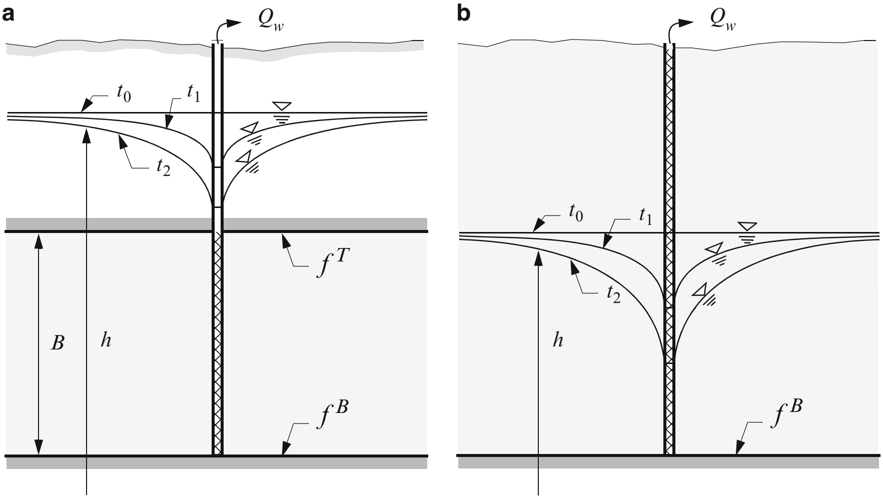


Fig. 9.11 Drawdown by a pumping well in (**a**) confined and (**b**) unconfined aquifer at elapsed times $t_0, t_1, t_2, \ldots$

with the well function

$$\begin{aligned} W(u) &= -\mathrm{Ei}(-u) = \int_u^{\infty} \frac{e^{-\xi}}{\xi} d\xi \\ &= -0.5772 - \ln(u) + \sum_{i=1}^{\infty} (-1)^{i-1} \frac{u^i}{i \cdot i!} \end{aligned} \tag{9.69}$$

and

$$u = \frac{r^s \bar{S}_o}{4Tt} \tag{9.70}$$

where r is the radial distance (coordinate) measured from the central well axis, h_0 is the initial piezometric head, $\bar{S}_o$ is the constant depth-integrated specific storage coefficient, T is the constant transmissivity, Q_w is the constant pumping rate and Ei() is the exponential integral.

For the numerical analysis of Theis' problem we study four different finite element schematizations as illustrated in Fig. 9.12:

- Mesh A: horizontal 2D structured discretization of a 30°-angle wedge configuration as symmetric part of the horizontally circular domain consisting of 1,135 linear quadrilateral elements with 1,368 nodes.
- Mesh B: horizontal 2D unstructured discretization of the complete circular domain consisting of 4,934 linear triangular elements with 2,506 nodes.

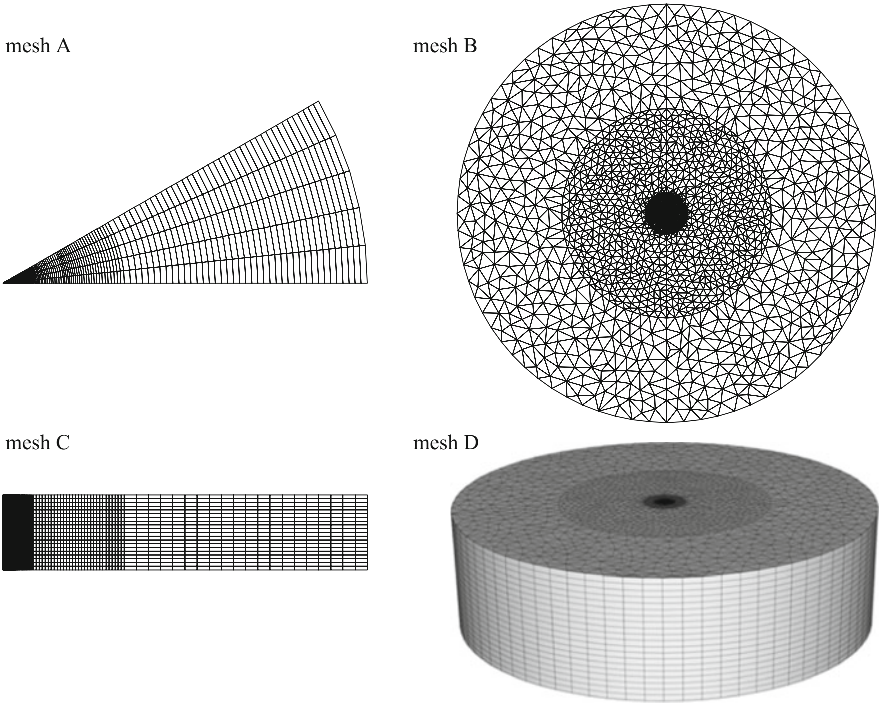


Fig. 9.12 Used finite element discretizations for simulating the Theis' problem: mesh A – 2D horizontal wedge configuration, mesh B – 2D horizontal mesh of the complete circular domain, mesh C – axisymmetric meridional cross-sectional schematization (vertical exaggeration 3:1) and mesh D – full 3D schematization (vertical exaggeration 10:1)

- Mesh C: vertical axisymmetric cross-sectional structured discretization formed by 3,200 linear quadrilaterals with 3,381 nodes.
- Mesh D: full 3D horizontally unstructured discretization formed by 20 mesh layers consisting of 98,680 linear pentahedral elements with 52,626 nodes in total.

The meshes have been suitably refined near the pumping well. We note that the problem formulation with meshes C and D does not need the Dupuit assumption. Table 9.1 summarizes the parameters used for the simulation. A comparison of the numerical results with the analytical Theis' solution is shown in Fig. 9.13 for the drawdown of the water table h in time at the radial distances $r = 2$ m and $r = 10$ m from the well and in Fig. 9.14 for the drawdown profiles at time $t = 10{,}000$ s. Up to a time of about 25,000 s the agreement of analytical and numerical results is quite well. Afterwards, the numerical solutions become influenced by the constant head BC at the outer boundary of the horizontally circular domain. Note that the Theis' solution assumes a constant head BC at infinite distance from the well. Clearly,

Table 9.1 Simulation parameters used for the Theis' problem

Quantity	Symbol	Value	Unit
Pumping well rate	Q_w	750	$\mathrm{m}^3\,\mathrm{d}^{-1}$
Wellbore radius	R	0.335	m
Outer boundary radius	R_Ω	300	m
Confined aquifer			
Aquifer thickness	B	20	m
Top of aquifer	f^T	20	m
Bottom of aquifer	f^B	0	m
Isotropic aquifer transmissivity	T	10^{-3}	$\mathrm{m}^2\,\mathrm{s}^{-1}$
Specific storage coefficient	$\bar{S}_o$	10^{-3}	1
Specific storage coefficient of well	S_{ow}	≈ 0	m^{-1}
Initial hydraulic head	h_0	30	m
Integral flux at recharge well	$\bar{q}_{n_h} = \frac{Q_w}{2\pi R}$	356.316	$\mathrm{m}^2\,\mathrm{d}^{-1}$
Flux rate at recharge well	$q_{n_h} = \frac{Q_w}{2\pi R B}$	17.816	$\mathrm{m}\,\mathrm{d}^{-1}$
Dirichlet-BC at outer boundary	$h_D = h(t, R_\Omega)$	30	m
Unconfined aquifer			
Bottom of aquifer	f^B	0	m
Isotropic hydraulic conductivity	K	$5 \cdot 10^{-5}$	$\mathrm{m}\,\mathrm{s}^{-1}$
Specific storage coefficient	S_o	$5 \cdot 10^{-5}$	m^{-1}
Specific storage coefficient of well	S_{ow}	≈ 0	m^{-1}
Specific yield	ε_e	0.2	1
Initial hydraulic head	h_0	20	m
Integral flux at recharge well	$\bar{q}_{n_h} = \frac{Q_w}{2\pi R}$	356.316	$\mathrm{m}^2\,\mathrm{d}^{-1}$
Dirichlet-BC at outer boundary	$h_D = h(t, R_\Omega)$	20	m
FEM			
Initial time step size	Δt_0	10^{-5}	d
Maximum error tolerance (AB/TR method)	ϵ	10^{-4}	1

to improve the drawdown curves for later times the computation domain must be enlarged.

The numerical results for the several meshes A, B, C and D differ only slightly and cannot be graphically distinguished in Figs. 9.13 and 9.14. Table 9.2 lists the actually obtained numerical values of the different meshes in comparison to the Theis' solution at time $t = 10{,}000$ s.

Due to the free surface BC in unconfined aquifer conditions the flow problem becomes nonlinear and there is no more an exact analytical solution (although some approximate solutions exist as discussed in [34]). In unconfined aquifers the temporal behavior of the drawdown is now significantly influenced by the specific yield, usually $\varepsilon_e \gg \bar{S}_o$, by which the process is much slower compared to confined aquifer conditions due to dewatering the unsaturated zone above the water table. On the other hand, the contraction of the water table in the vicinity of a pumping well is associated with vertical flow components which can be significant so the Dupuit assumption of horizontal flow is no more applicable, at least in the near-field

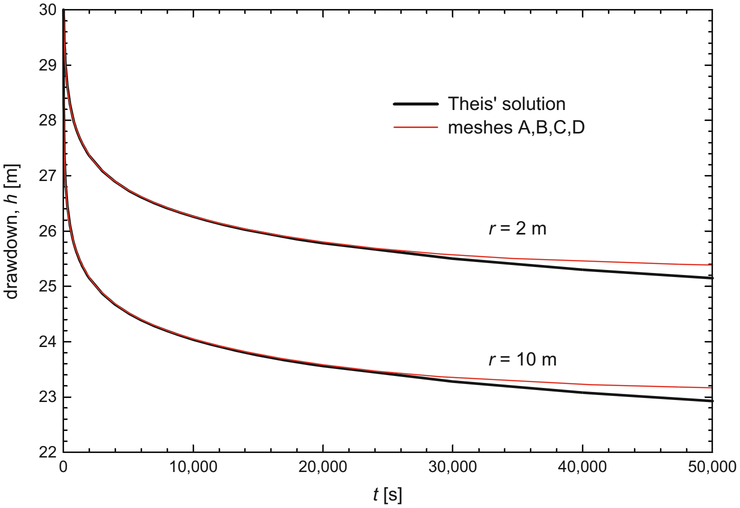


Fig. 9.13 Analytically versus numerically computed drawdown of water table h in time t at radial distances $r = 2\,\text{m}$ and $r = 10\,\text{m}$ from the well for confined aquifer condition

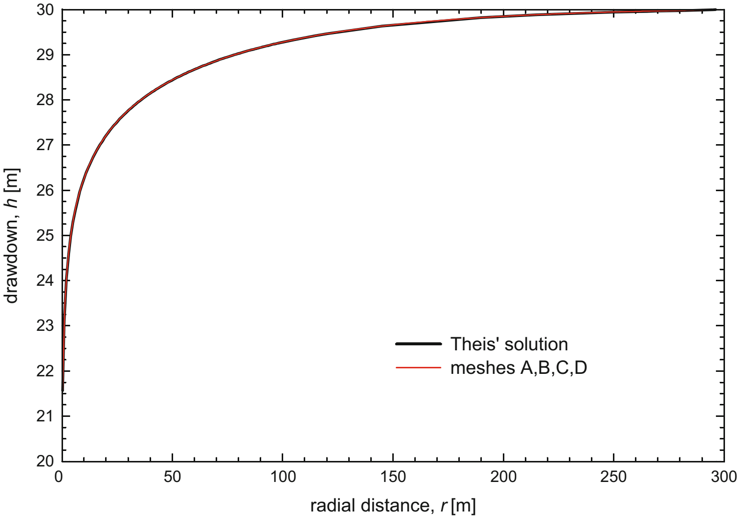


Fig. 9.14 Computed profile of water table h at time $t = 10{,}000\,\text{s}$ along the radial distance r from the well for confined aquifer condition

of the well. We perform the computations for the unconfined aquifer with the data of Table 9.1. Simulation results obtained by the 2D horizontal Dupuit-based models are compared with 3D model results in Fig. 9.15. For the 3D models both

Table 9.2 Drawdown h (m) computed for meshes A, B, C and D at time $t = 10{,}000\,\text{s}$ in comparison to the Theis' solution

Distance r (m)	Theis' solution	Mesh A	Mesh B	Mesh C	Mesh D
2	24.0364	24.0468	24.0591	24.0389	24.0583
10	26.2582	26.2655	26.2737	26.1613	26.2730

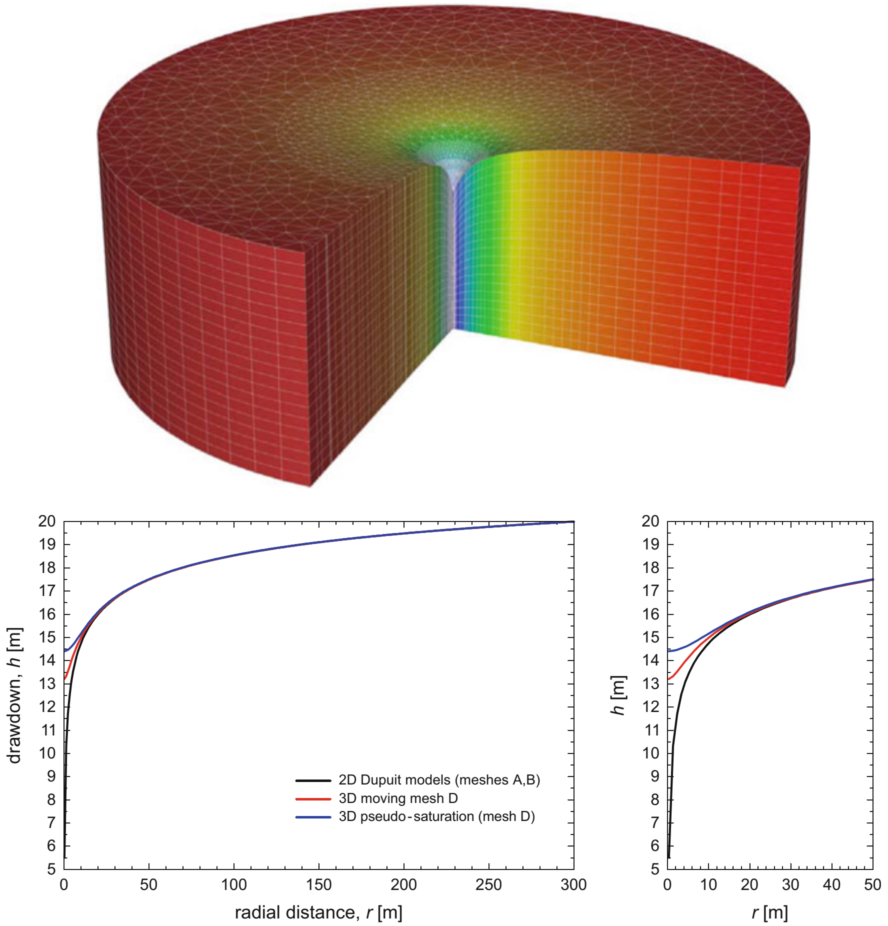


Fig. 9.15 Cut view of moving mesh D (vertical exaggeration 10:1) for free-surface modeling and computed profiles of water table h at time $t = 100\,\text{d}$ along the radial distance r from the well for unconfined aquifer condition

moving mesh with BASD technique (see Fig. 9.15 for mesh D) and fixed mesh with pseudo-saturation strategies are applied. The comparison in form of the free-surface profiles in Fig. 9.15 reveals a reasonable agreement in larger distances from the well, however, near to the well the water table differs. We observe that Dupuit-based models give a higher drawdown at the well while 3D modeling approaches do produce an obviously more realistic inflow field to the well. It is to be noted that the well flux condition is realized in 3D models via a multilayer well BC (Sect. 6.5.6) and in 2D horizontal meshes via an integral Neumann-type flux BC (Sect. 6.5.4).

9.8.2 3D Anisotropy and Flow Patterns of Groundwater Whirls

3D anisotropy in layered aquifers can lead to somewhat unusual flow patterns, even in steady-state flow situations. This was firstly discovered and reported by Hemker et al. [239, 241], who termed this type of flow pattern as groundwater whirls. The cause and appearance of groundwater whirls have been thoroughly discussed and studied using both numerical and analytical solution techniques [22, 240, 370].

9.8.2.1 Two-Layer Crosswise Anisotropy

Hemker et al. [241] considered a box-shaped, two-layer, homogeneous aquifer. The model box is 200 m long, 70 m wide, and 20 m thick, and consists of only two layers, each 10 m thick. An anisotropic block located at the center of the box is 150 m long, 20 m wide, and 20 m thick. Inside the anisotropic block, the major principal directions of the horizontal hydraulic conductivity tensor are orthogonal. The general flow direction is lengthwise (straight north) and makes an angle of 45° with either of these directions (see Fig. 9.16).

The hydraulic conductivity of the isotropic outer area is $1\,\mathrm{m\,d^{-1}}$. Within the anisotropic block, the major principal value of the horizontal conductivity tensor is $K_{\max} = 1\ \mathrm{m\,d^{-1}}$, and the minor principal value is $K_{\min} = 0.1\,\mathrm{m\,d^{-1}}$, so that the anisotropy ratio is $\Xi_{\mathrm{aniso}} = K_{\min}/K_{\max} = 0.1$. The major principal directions of the horizontal hydraulic conductivity in the two layers are perpendicular to each other – southeast-northwest in the upper layer, and southwhest-northeast in the lower layer, which is referred as a crosswise anisotropy. The vertical hydraulic conductivity in the isotropic and anisotropic parts is $1\,\mathrm{m\,d^{-1}}$.

The flow in the aquifer is fully confined. The western and eastern sides are no-flow boundaries, while the short southern and northern sides are open boundaries with fixed hydraulic heads that differ by 1 m. It is to be noted that the specific value of the gradient has an effect on the values of the head contours and velocity only, and does not affect the pattern of head contours and flow lines [241].

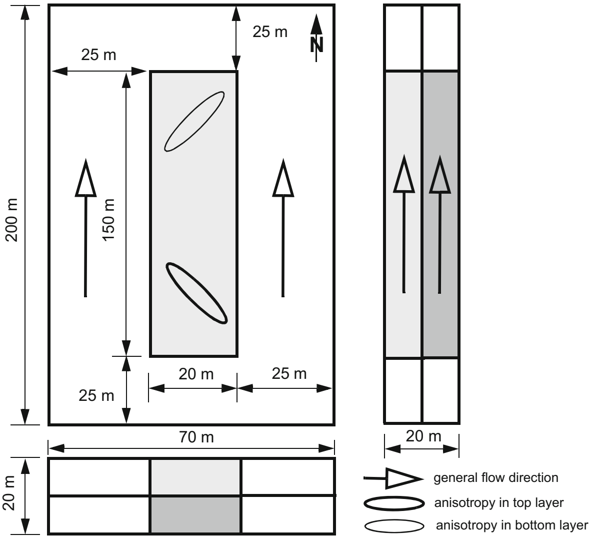


Fig. 9.16 The two-layer anisotropic model problem (Modified from [241])

We use a 3D mesh consisting of hexahedral brick-shaped elements, where the vertical cross section of the two-layer aquifer problem coincides with the $x_1 - x_2$−plane, while the x_3−coordinate coincides with the horizontal north direction. For such a coordinate orientation the Eulerian anisotropic angles have to be defined for the general rotation matrix $\boldsymbol{A}$ of anisotropy according to (7.12), viz.,

$$\boldsymbol{A} = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \quad \text{at} \quad \begin{aligned} \phi &= -90^\circ \\ \psi &= \ 90^\circ \end{aligned} \tag{9.71}$$

where the rotation is performed about the x_2−axis by the angle θ as the only anisotropic angle to be input for the present problem at each layer of the inner area of the model domain. The mesh used in FEFLOW is shown in Fig. 9.17. It consists in total of 640,000 hexahedral elements with 667,521 nodes. In the cross-sectional $x_1 - x_2$−plane the nodal distances are 1.25 m in the x_1−direction and 0.5 m in the x_2−direction for the isotropic outer area and 0.5 m in both directions for the anisotropic inner area. The nodal distance along the horizontal x_3−direction is chosen always 1 m. The parameters used for the simulation are summarized in Table 9.3.

We compare the FEFLOW results with the outcome from Hemker et al. [241]. Five flow pathlines are shown in Fig. 9.18, all starting in the upper layer at the center of the southern model boundary, at depths of 1, 3, 5, 7, and 9 m from the confined top; the pathlines are numbered 1 through 5 from top to bottom. Isochrone markers on the pathlines indicate a residence time interval of 10 years, based on

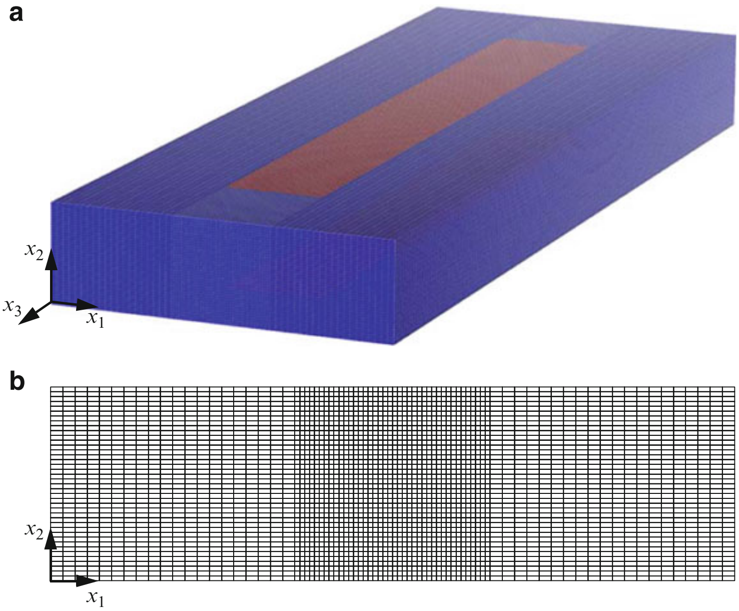


Fig. 9.17 FEFLOW mesh used for simulating the two-layer anisotropy model problem consisting of 640,000 hexahedral elements with 667,521 nodes: (**a**) 3D view, (**b**) cross-sectional view

30 % porosity. The computed total travel times for the five pathlines are compared in Table 9.4 with the results of Hemker et al. [241]. Figure 9.19 displays the pathlines in a 3D view, which clearly indicate a typical whirl pattern of the flow field.

There is a reasonable quantitative agreement between Hemker et al.'s and the present FEFLOW results, however, some differences exist in the pathline characteristics (Fig. 9.18) and the computed travel times (Table 9.4), which are obviously caused by the different mesh resolutions used (note that the FEFLOW mesh is more resolved that Hemker et al.'s mesh). Particularly, pathlines 3, 4, and 5 become increasingly depart from Hemker et al.'s results with progressing travel time. This has to be expected because these pathlines start at decreasing distance from the interface of the two layers, where the influence of the strong crosswise anisotropy becomes more dominant and consequently a higher numerical accuracy should be required there.

Hemker et al. [241] reported a total discharge of $6.3\,\mathrm{m}^3\,\mathrm{d}^{-1}$. The FEFLOW simulation result a discharge of $6.18\,\mathrm{m}^3\,\mathrm{d}^{-1}$. Water balance computations for each layer show that vertical flow components between both layers are induced. We find an exchanging rate of $3.2\,\mathrm{m}^3\,\mathrm{d}^{-1}$ in both directions at the interface of the two layers, which is identical to Hemker et al.'s result.

Table 9.3 Simulation parameters used for the two-layer anisotropy model problem

Quantity	Symbol	Value	Unit
Length of outer area		200	m
Length of inner area		150	m
Width of outer area		70	m
Width of inner area		20	m
Thickness of each layer		10	m
Steady-state flow			
Potential difference in north direction	Δh	1	m
Impervious boundaries, except for vertical front and back faces			
Isotropic conductivity of outer area	$K_1^m = K_2^m = K_3^m$	$1.15741 \cdot 10^{-5}$	$\mathrm{m\,s^{-1}}$
Major principal conductivity of inner area	$K_1^m = K_2^m = K_{\max}$	$1.15741 \cdot 10^{-5}$	$\mathrm{m\,s^{-1}}$
Minor principal conductivity of inner area	$K_3^m = K_{\min}$	$1.15741 \cdot 10^{-6}$	$\mathrm{m\,s^{-1}}$
Eulerian angles for outer area (all two layers)	$\phi = \theta = \psi$	0	°
Eulerian angle for inner area (all two layers)	ϕ	−90	°
Eulerian angle for inner area (all two layers)	ψ	90	°
Eulerian angle for inner area (only upper layer)	θ	45	°
Eulerian angle for inner area (only lower layer)	θ	−45	°

9.8.2.2 Model of Nine-Layer Randomly Distributed Anisotropy

Hemker and Bakker [239] studied a 18 m thick confined prototypical aquifer at steady state, consisting of nine equally thick layers. The horizontal hydraulic conductivity is heterogeneous in a 100 m wide section of the model; each layer in this section is divided in 10 strips of equal width. A cross section perpendicular to these strips shows a regular pattern of 9-by-10 cells, where each cell is 10 m wide and 2 m high (Fig. 9.20). On each side of this central zone a 100 m wide homogeneous block serves to reduce boundary effects.

The major and minor principal values of the horizontal hydraulic conductivity tensor are 10 and 5 $\mathrm{m\,d^{-1}}$, respectively, in the entire model. The vertical hydraulic conductivity is 1 $\mathrm{m\,d^{-1}}$ in all layers. The general flow direction is in the direction of the strips (straight north). In the presented model the major principal direction of the horizontal hydraulic conductivity tensor is also chosen straight north in the two large side blocks, while it varies between −45° (N45W, northwest) and 45° (N45E, northeast) in the 90 cells of the central zone. To obtain a 2D spatial distribution, ten

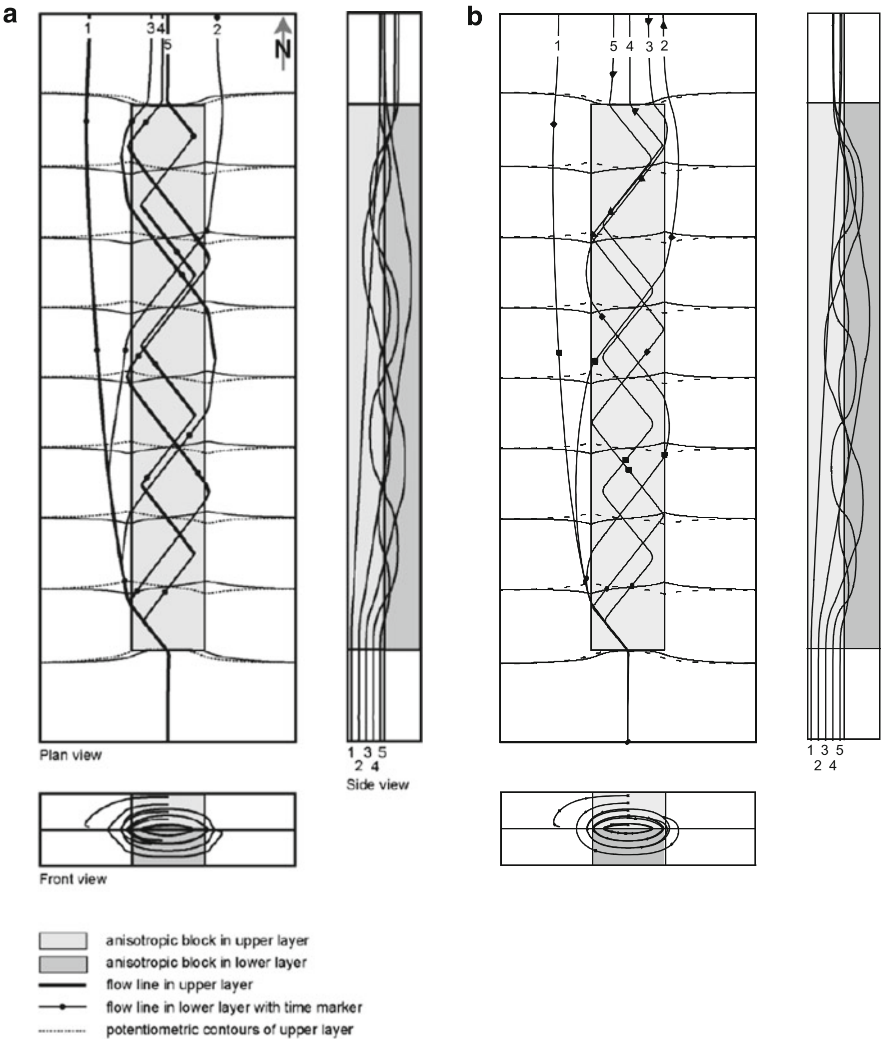

a
1
34
5
2
N
Plan view
1 3 5
2 4
Side view
Front view
b
1
5 4 3 2
1 3 5
2 4
anisotropic block in upper layer
anisotropic block in lower layer
flow line in upper layer
flow line in lower layer with time marker
potentiometric contours of upper layer

Fig. 9.18 Plan, side, and front view of the two-layer model problem with hydraulic head contours of both layers and five pathlines (1, 2, 3, 4, 5) starting at different levels in the upper layer on the southern boundary: (**a**) results by Hemker et al. [241], (**b**) FEFLOW results

uniformly distributed anisotropy directions (α−angle, see Fig. 9.20) were chosen ($-45°, -35°, -25°, \ldots, 45°$) and for each layer these ten values were randomly assigned to the cells. The resulting distribution is given in Table 9.5. The model size is 300 by 300 m by 18 m. The east and west sides are no-flow boundaries, while the south and north sides are open boundaries with fixed hydraulic heads that differ by 0.3 m in all layers.

For the present simulations we choose a similar mesh resolution comparable to Hemker and Bakker [239]. It is convenient to use again the vertical cross section as

Table 9.4 Comparison of the computed travel times for the five pathlines of the two-layer model problem

	Travel times (years)	
Pathline no.	Hemker et al. [241]	FEFLOW
1	35.2	35.4
2	40.2	40.6
3	46.7	50.4
4	56.6	55.6
5	57.7	53.2

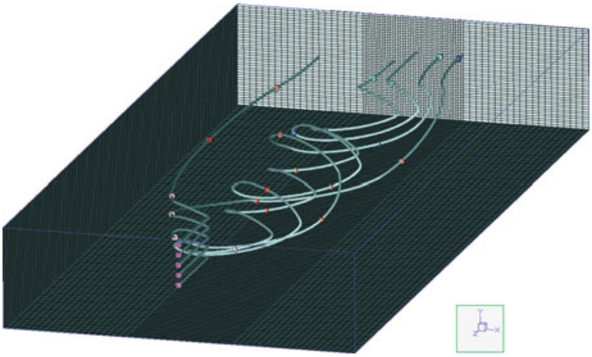


Fig. 9.19 Whirling flow pattern appearing in the 3D view of the computed flow pathlines (FEFLOW results)

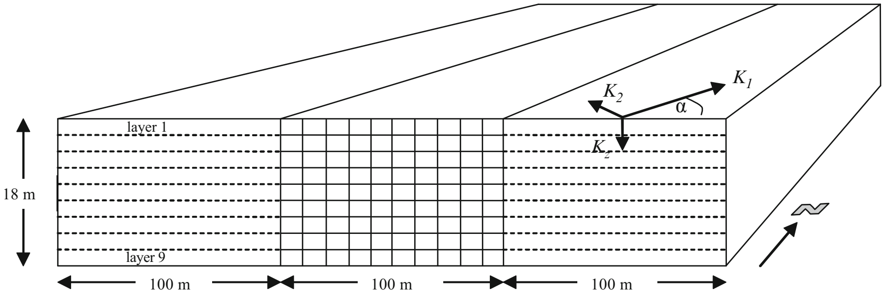


Fig. 9.20 A stratified confined aquifer with a laterally heterogeneous anisotropic central zone (Taken from [239])

the x_1-x_2-plane so that the horizontal extent of the aquifer system is oriented to the x_3-coordinate (north) direction. According to Table 9.5 the anisotropic angle is only defined around the vertical axis. In using such a coordinate alignment it means – in the same manner as for the preceding example – that the rotation is done explicitly via the Eulerian angle θ for the rotation around the vertical x_2-axis, cf. (9.71).

Table 9.5 Principal directions of horizontal anisotropy in all 9 by 12 cells of the model (angle α in ° as defined in Fig. 9.20)

0	−35	5	15	35	−15	−25	−5	25	−45	45	0
0	25	35	−35	15	−45	−5	−25	5	−15	45	0
0	−35	25	−15	−25	−5	15	5	−45	45	35	0
0	−35	35	5	−25	15	−5	45	25	−15	−45	0
0	15	−5	25	−45	35	45	5	−35	−15	−25	0
0	−45	−35	25	−25	5	−15	15	35	−5	45	0
0	25	−45	−5	35	−25	5	15	−15	45	−35	0
0	35	5	−15	−35	25	−45	−25	15	45	−5	0
0	45	−25	−35	−5	−15	15	5	−45	35	25	0

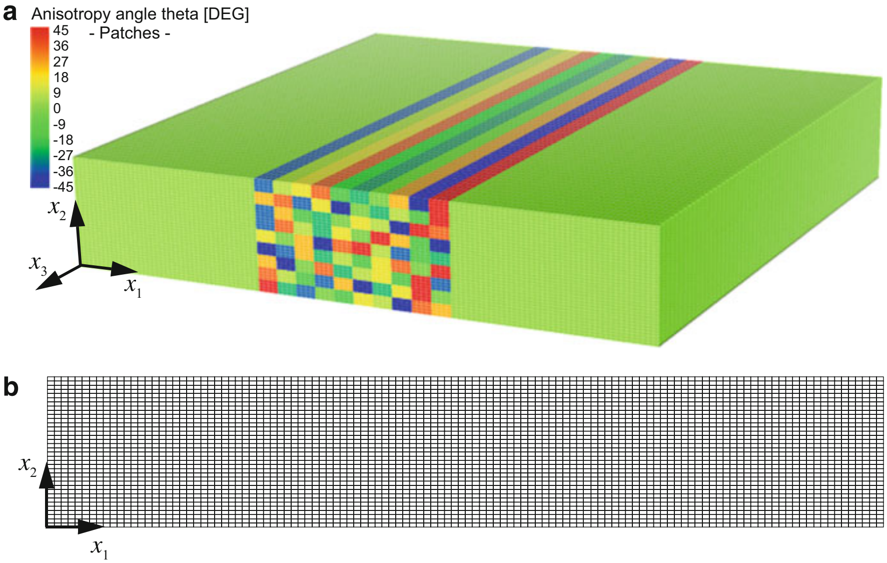


Fig. 9.21 FEFLOW mesh used for the nine-layer anisotropy model problem consisting of 259,200 hexahedral elements with 273,097 nodes: (**a**) 3D view with drawn $\theta-$angle pattern of anisotropy, (**b**) cross-sectional view (vertical exaggeration 3:1)

The used FEFLOW mesh is shown in Fig. 9.21. It consists of 259,200 hexahedral elements with 273,097 nodes. In the cross-sectional $x_1 - x_2-$plane the nodal distances are 2.5 m in the x_1-direction and 0.5 m in the x_2-direction. The nodal distance along the horizontal x_3-direction is constantly 5 m. The parameters used for the simulation are summarized in Table 9.6.

Due to the anisotropy of the layered aquifer system a complex whirling flow pattern results. Large and small whirls exist next to each other, rotating in opposite directions. A comparison of the FEFLOW results with the numerical and analytical results presented by Hemker and Bakker [239] reveal a rather good agreement. Figure 9.22 displays the computed whirl pattern at a vertical cross section.

Table 9.6 Simulation parameters used for the nine-layer anisotropy model problem

Quantity	Symbol	Value	Unit
Domain measure (length; width; height)		300; 300; 18	m
Width of inner zone		100	m
Width of outer zone (left and right)		100	m
Nine layers with thickness of each layer		2	m
Steady-state flow			
Potential difference in north direction	Δh	0.3	m
Impervious boundaries, except for vertical front and back faces			
Conductivity for all elements	K_1^m	$1.15741 \cdot 10^{-4}$	$\mathrm{m\,s^{-1}}$
Conductivity for all elements	K_2^m	$1.15741 \cdot 10^{-5}$	$\mathrm{m\,s^{-1}}$
Conductivity for all elements	K_3^m	$5.78704 \cdot 10^{-5}$	$\mathrm{m\,s^{-1}}$
Eulerian angle for all elements	ϕ	-90	°
Eulerian angle for all elements	ψ	90	°
Eulerian angle for all elements of isotropic outer zone	θ	0	°
Eulerian angle for all elements of anisotropic inner zone	θ	α−angles of Table 9.5	

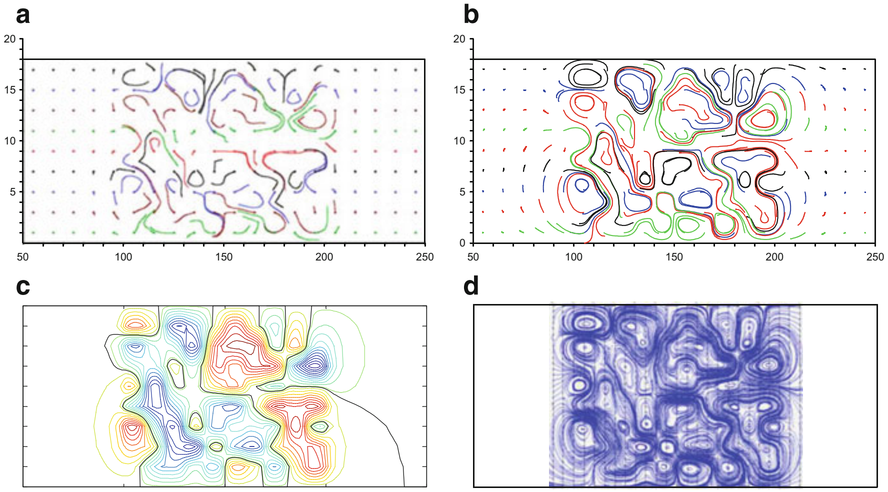


Fig. 9.22 Cross section showing the pattern of groundwater whirls: (**a**) Hemker and Bakker's results [239]. Projection of 20 pathlines in each of the 9 layers, pathlines start in the center of each cell and at equal distances to the left and right in the homogeneous side blocks. All starting points are located at 100 m north of the southern boundary and all pathlines run for 100 m due north. Different colors are used for different starting depths. (**b**) FEFLOW results using the same starting points at 100 m north of the southern boundary, however, pathlines run for 200 m due north. (**c**) Analytical results of the streamfunction contours in the cross section. (**d**) FEFLOW results for a refined pathline pattern at the cross section, pathlines run over the full horizontal distance (300 m) due north

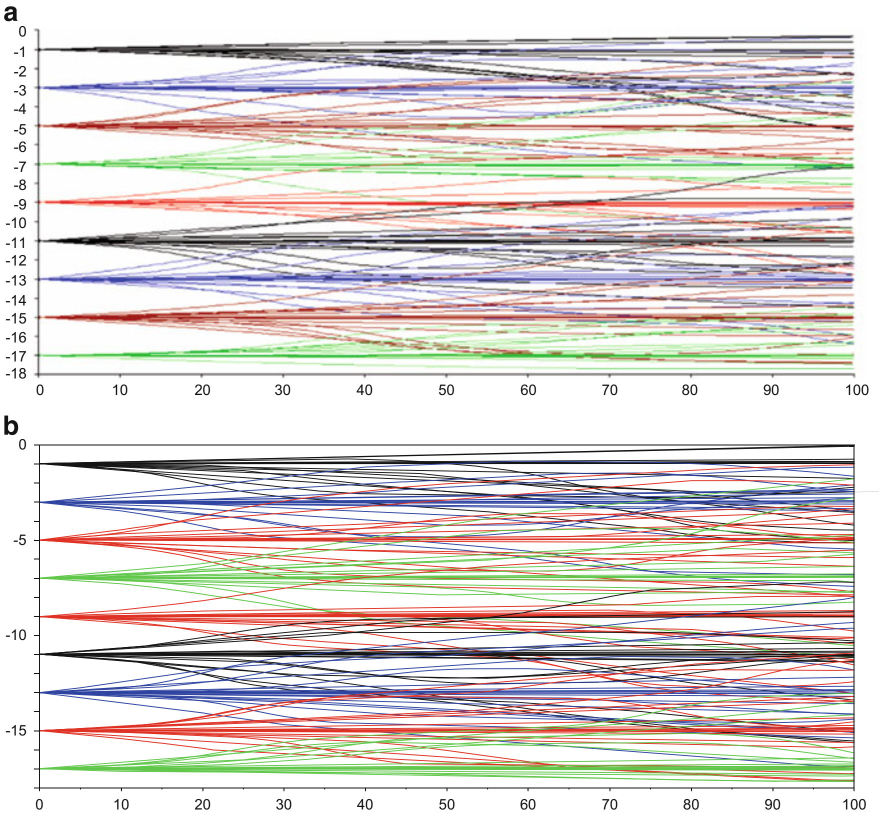


Fig. 9.23 The 9 × 20 pathlines projected on a side view: (**a**) Hemker and Bakker's results [239] (**b**) FEFLOW results

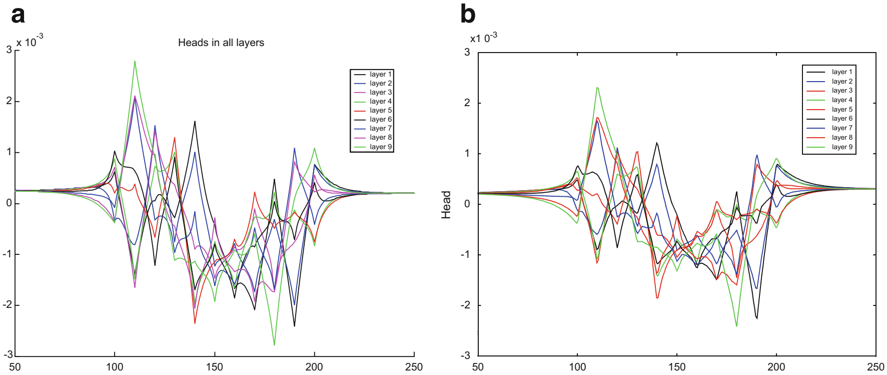


Fig. 9.24 Hydraulic heads in an east-west cross section of the layered aquifer: (**a**) analytical profiles [239], (**b**) present FEFLOW results

Evidently, the numerical results agree very well with the analytical predictions. The same pathlines projected on a side view are depicted in Fig. 9.23 comparing the FEFLOW results with the Hemker and Bakker's results [239]. Slight differences in the divergent characteristics of the pathlines can be recognized that are obviously caused by the different meshes, velocity computations and pathline tracking techniques used in the different codes. Comparing FEFLOW's hydraulic head distribution along a cross section with the analytical (exact) results presented by Hemker and Bakker [239] we found in Fig. 9.24 that while the profiles agree rather well, the peaks in the zigzag-profiles of the hydraulic head appear generally somewhat smaller in the numerical predictions. More refined 3D meshes seem to be needed to improve the agreement with the analytical results.

Chapter 10
Flow in Variably Saturated Porous Media

10.1 General

Flow processes in variably saturated porous media are subject in many disciplines. Most notably are applications in soil sciences and subsurface hydrology, for instance the study of infiltration processes for analyzing the movement and spreading of water from the ground surface to the water table. However, it is also an increasingly significant interest in other fields such as petroleum and geotechnical engineering, material research for industrial porous media and many others. From the physical point of view, flow in variably saturated porous media represents a generalization of porous-media flow in which fully saturated media are a special case. On the other hand, there is a clear distinction due to presence of capillarity which introduces a new physical quality and gives rise to completely different flow situations. Most remarkable are flows which occur in isolated regions (e.g., perched water or fingered flow) or flow conditions forming capillary barrier effects where flow will not cross between zones of differing properties. Free surface conditions in phreatic aquifers can be revealed in a more general context without the need for assuming a coherent interface between unsaturated and saturated zones as required in free-surface groundwater flow modeling of Chap. 9.

Traditionally, it is often distinguished between unsaturated and saturated porous media. In the context of subsurface hydrology, the unsaturated zone is sometimes termed as vadose zone defining the range between the ground surface and the water table. An unsaturated medium is defined as a zone where the liquid phase (usually water) has a saturation less than 100 % and the gas phase (usually air) is assumed to be stagnant, see definitions and derivations introduced in Sects. 3.8.7 and 3.10 as well as Appendix D. In physical terms the differences between unsaturated media and saturated media are listed in Table 10.1. The negative (capillary) pressure head ψ and the related saturation s less than unity represent the striking features of an unsaturated porous medium. An unsaturated medium requires additional relations in form of retention $\psi = \psi(s)$ and relative permeability $k_r = k_r(s)$, which have to be specified by empirical parametric expressions fitted from experimental data

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_10,

Table 10.1 Unsaturated versus saturated porous-media conditions

Unsaturated media	Saturated media
$\psi < 0,\ \psi = \psi(s),\ s_r < s < 1$	$\psi \geq 0,\ s = 1$
$k_r = k_r(s)$	$k_r = 1$

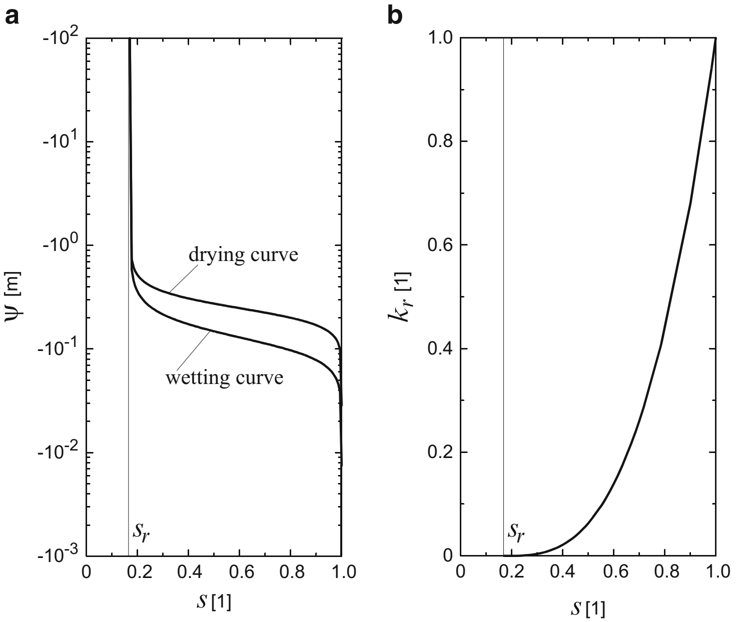


Fig. 10.1 Typical plots of (**a**) retention curve $\psi(s)$ with hysteretic behavior and (**b**) relative permeability $k_r(s)$ curve

(see Appendix D). Strong dependencies on the saturation s are typical (Fig. 10.1) which make the unsaturated flow processes heavily nonlinear. It is generally not useful to differ between unsaturated and saturated flow a priori. The saturation of a porous medium represents a dynamic state variable which can vary both spatially and temporarily. Accordingly, solution strategies are required which are capable of computing *simultaneous unsaturated-saturated flow*. In this way, we prefer a general, accurate and efficient modeling approach which handles the full spectrum of pressure head $-\infty < \psi < \infty$ without any significant restrictions for flow in variably saturated and heterogeneous porous systems.

10.2 Basic Equations

The basic equations for 3D and 2D (including axisymmetric) flow in variably saturated porous media have been developed in Sect. 3.10.5 and can be taken from Table 3.7. They are

$$\begin{aligned} s\,S_o\frac{\partial h}{\partial t}+\varepsilon\frac{\partial s}{\partial t}+\nabla\cdot\boldsymbol{q} &= Q+Q_{\mathrm{EOB}} \\ \boldsymbol{q} &= -k_r\boldsymbol{K}\,f_\mu\cdot\big(\nabla h+\chi\boldsymbol{e}\big) \end{aligned} \tag{10.1}$$

written in terms of the hydraulic head variable $h = \psi + z$, or alternatively

$$\begin{aligned} s\,S_o\frac{\partial \psi}{\partial t}+\varepsilon\frac{\partial s}{\partial t}+\nabla\cdot\boldsymbol{q} &= Q+Q_{\mathrm{EOB}} \\ \boldsymbol{q} &= -k_r\boldsymbol{K}\,f_\mu\cdot\big[\nabla \psi+(1+\chi)\boldsymbol{e}\big] \end{aligned} \tag{10.2}$$

written in terms of the pressure head variable ψ, where the following dependencies exist

$$s = s(\psi), \quad k_r = k_r(s) \tag{10.3}$$

which have to be specified in form of constitutive relationships as summarized in Appendix D. Most known are the empirical analytic relationships proposed by van Genuchten (D.4), (D.26) and Brooks-Corey (D.8), (D.30). Other useful relations are discussed in Appendix D, including spline approximations for $s(\psi)$ and $k_r(s)$. Their inherent parameters have to be fitted from measured data. The retention curve $s(\psi)$ can be used to convert the saturation variable s to the pressure head variable ψ (or hydraulic head variable $h = \psi + z$). On the other hand, we can also assume that the retention curve $s(\psi)$ is uniquely invertible so that

$$s = f(\psi), \quad \psi = f^{-1}(s) \tag{10.4}$$

which is exemplified by (D.4) and (D.5) for the analytic van Genuchten retention curve.

Usually, $\boldsymbol{q}$ is substituted by the Darcy equation to obtain the governing Richards' equation (cf. Sect. 3.11) written in the $h-s-$form and $\psi-s-$form, respectively,

$$\begin{aligned} s\,S_o\frac{\partial h}{\partial t}+\varepsilon\frac{\partial s}{\partial t}-\nabla\cdot\big[k_r\boldsymbol{K}\,f_\mu\cdot(\nabla h+\chi\boldsymbol{e})\big] &= Q_h+Q_{hw}+Q_{\mathrm{EOB}} \\ s\,S_o\frac{\partial \psi}{\partial t}+\varepsilon\frac{\partial s}{\partial t}-\nabla\cdot\big[k_r\boldsymbol{K}\,f_\mu\cdot(\nabla \psi+(1+\chi)\boldsymbol{e})\big] &= Q_h+Q_{hw}+Q_{\mathrm{EOB}} \end{aligned} \tag{10.5}$$

where the source/sink term $Q = Q_h + Q_{hw}$ is suitably split into a supply term Q_h and a well-type SPC term Q_{hw}. The PDE's (10.5) have to be solved for the remaining variables h, ψ and/or s subject to a set of BC's of Dirichlet, Neumann, gradient and Cauchy type as well as well-type SPC as introduced in Sects. 6.3.1 and 6.5.5, written in the h variable:

$$\begin{aligned} h &= h_D && \text{on} \quad \varGamma_D\times t[t_0,\infty) \\ -[k_r\boldsymbol{K}\,f_\mu\cdot(\nabla h+\chi\boldsymbol{e})]\cdot\boldsymbol{n} &= q_h && \text{on} \quad \varGamma_N\times t[t_0,\infty) \end{aligned}$$

$$\begin{aligned} -[\boldsymbol{K} f_\mu \cdot (1+\chi)\boldsymbol{e})]\cdot\boldsymbol{n} &= q_h^{\nabla} && \text{on} \quad \varGamma_N^{\nabla}\times t[t_0,\infty)\\ -[k_r\boldsymbol{K} f_\mu \cdot (\nabla h+\chi\boldsymbol{e})]\cdot\boldsymbol{n} &= -\varPhi_h(h_C-h) && \text{on} \quad \varGamma_C\times t[t_0,\infty)\\ Q_{hw} &= -\sum_w Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w\in\varOmega\times t[t_0,\infty) \end{aligned} \tag{10.6}$$

where on $\varGamma_N^{\nabla}$ the special Neumann BC in form of the gradient-type BC is specified, cf. (6.76), also the nonlinear BC of a seepage face as introduced in Sect. 6.5.2,

$$\left.\begin{array}{r} h = z \\ Q_{n_h} < Q_{n_h}^{\max_1} = 0 \end{array}\right\} \quad \text{on} \quad \varGamma_S\times t[t_0,\infty) \tag{10.7}$$

and in combination with the IC of the form

$$h(\boldsymbol{x},t_0) = h_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{10.8}$$

where the total boundary is $\varGamma = \varGamma_D \cup \varGamma_N \cup \varGamma_N^{\nabla} \cup \varGamma_C \cup \varGamma_S$. Alternative BC and IC formulations can be expressed via the ψ and s variables if using the relations $h = \psi + z$ and $\psi = f^{-1}(s)$, e.g., for prescribing IC's of the pressure head ψ_0 and saturation s_0:

$$h(\boldsymbol{x},t_0) = \begin{cases} \psi_0(\boldsymbol{x}) + z & \text{or} \\ f^{-1}\big(s_0(\boldsymbol{x})\big) + z \end{cases} \quad \text{in} \quad \bar{\varOmega} \tag{10.9}$$

Once (10.5) has been solved, the secondary variable of Darcy velocity $\boldsymbol{q}$ can be evaluated as a derived quantity of known h or ψ. The essential parameters required for solving (10.5) with (10.6)–(10.8) (or (10.9)) are listed in Tables I.1 to I.6 and I.10 of Appendix I in accordance with the chosen problem type. *Steady-state* flow situations can only occur under unsaturated conditions, $s < 1$ with $\varepsilon > 0$, if $\partial s/\partial t$ (and accordingly $\partial h/\partial t$) approaches to zero and under saturated conditions with $s = 1$ if $S_o = 0$ (incompressibility) or $\partial h/\partial t$ approaches to zero.[1]

10.3 Three Essential Forms of the Richards' Equation and Choice of Primary Variables

Obviously, the governing flow equation (10.5) for the flow in variably saturated porous media involves two essential solution variables in form of h (or synonymously ψ) and s. Since only one flow equation is available it is to be decided which

[1] Optionally, FEFLOW suppresses the time derivative terms $\partial s/\partial t$ and $\partial h/\partial t$ for solving steady-state solutions.

variable is primary and which must be secondary. Depending on such a choice, different modeling approaches result which are mathematically equivalent in the continuous formulation, but their discrete analogs are different providing their own advantages and drawbacks:

1. The natural form of the Richards' equation as derived above with (10.5) is the *mixed* $\psi - s-$form (or the equivalent $h - s-$form), where both variables are employed and, in solving the resulting discrete equation system, the pressure head ψ (or hydraulic head h) is usually used as the primary variable. However, it will be seen further below, there is also a solution strategy based on the mixed $\psi - s-$form of the Richards' equation in which the primary variable can be dynamically switched between ψ and s in accordance with the computational requirements. Numerical schemes based on the mixed form possess superior properties with respect to accurate mass conservation solutions, e.g., [72, 141, 361]. The mixed-form Richards' equation is applicable to both unsaturated and saturated conditions. The hydraulic head (or pressure head) variable is unique and continuous, regardless of whether the porous medium is homogeneous or heterogeneous.
2. The *standard* Richards' equation represents a $\psi-$based (or the equivalent $h-$based) form if the 1st derivative of saturation appearing in (10.5) is expressed by

$$\frac{\partial s}{\partial t} = \frac{\partial s}{\partial \psi}\frac{\partial \psi}{\partial t} = C\,\frac{\partial \psi}{\partial t}, \quad \frac{\partial s}{\partial t} = \frac{\partial s}{\partial \psi}\frac{\partial \psi}{\partial h}\frac{\partial h}{\partial t} = C\,\frac{\partial h}{\partial t} \tag{10.10}$$

where

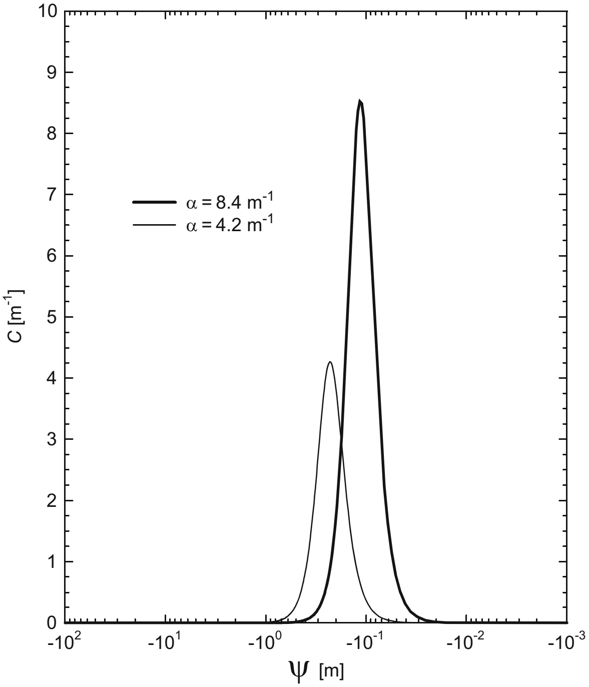
$$C = \frac{\partial s}{\partial \psi} \tag{10.11}$$

is the *moisture capacity* which can be derived from given retention relation $s = s(\psi)$ provided $s(\psi)$ is continuously differentiable at any ψ. Most important $C-$relationships are summarized in Appendix D. Typical graphs of $C = C(\psi)$ are illustrated in Fig. 10.2 for the analytic van Genuchten relationship which reveals a monotonic behavior. Inserting (10.10) into (10.5) the $\psi-$form (or the equivalent $h-$form) of the Richards' equation results

$$\begin{aligned} (s\,S_o + \varepsilon C)\frac{\partial \psi}{\partial t} - \nabla\cdot\big[k_r \boldsymbol{K} f_\mu \cdot (\nabla\psi + (1+\chi)\boldsymbol{e})\big] &= \quad Q_h + Q_{hw} + Q_{\text{EOB}} \\ (s\,S_o + \varepsilon C)\frac{\partial h}{\partial t} - \nabla\cdot\big[k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi\boldsymbol{e})\big] &= \quad Q_h + Q_{hw} + Q_{\text{EOB}} \end{aligned} \tag{10.12}$$

to solve ψ or h. The $\psi-$form (and the equivalent $h-$form) of the Richards' equation is applicable to both unsaturated and saturated conditions. The pressure head (or hydraulic head) variable is unique and continuous. Models of this type

Fig. 10.2 Typical plots of moisture capacity $C(\psi) = \partial s/\partial\psi$ of the van Genuchten relationship (D.6) for different parameters α (at $n = 5.25$, $m = 1 - 1/n$, $s_r = 0.17$ and $s_s = 1$)



have been extensively used. But, it has been shown, e.g., [72, 141, 373], that the $\psi-$based (and the equivalent $h-$based) form can produce significant global mass balance errors under certain saturation conditions unless very small time steps are used in the numerical approach. The mass balance errors are caused in approximating the storage term $\partial s/\partial t$ by the expansion $C(\psi)\partial\psi/\partial t$ in a discrete manner. A certain remedy can be attained if the moisture capacity term $C(\psi)$ is performed by suited chord slope approximations in replacing direct analytic derivatives as discussed further below.

3. It is possible to express completely the Richards' equations (10.5) in terms of the $s-$variable. This can be done by invoking the derivatives

$$\frac{\partial\psi}{\partial t} = \frac{\partial\psi}{\partial s}\frac{\partial s}{\partial t} = C^{-1}\frac{\partial s}{\partial t}, \quad \nabla\psi = \frac{\partial\psi}{\partial s}\nabla s = C^{-1}\nabla s \tag{10.13}$$

to obtain a common $s-$form of the Richards' equation

$$\big(s\,S_o\,C^{-1} + \varepsilon\big)\frac{\partial s}{\partial t} - \nabla\cdot\big[\boldsymbol{D}\cdot\nabla s + k_r\boldsymbol{K}f_\mu(1+\chi)\boldsymbol{e}\big] = Q_h + Q_{hw} + Q_{\text{EOB}} \tag{10.14}$$

for solving the saturation s, where

$$\boldsymbol{D} = k_r\boldsymbol{K}f_\mu C^{-1} \tag{10.15}$$

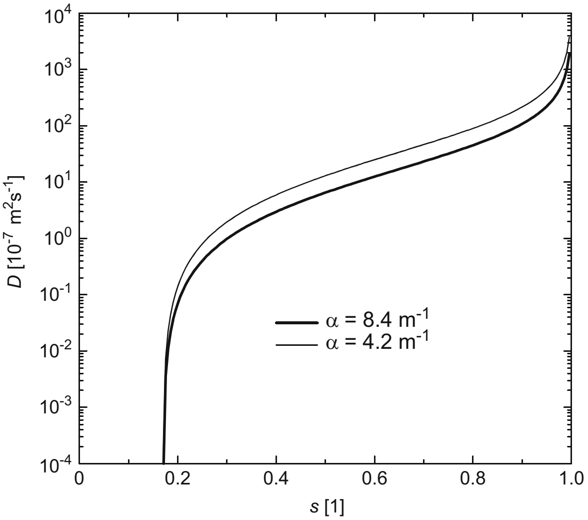


Fig. 10.3 Typical plots of (scalar) capillary diffusivity $D = k_r K C^{-1}$ using the van Genuchten relationships (D.6) and (D.26) for different parameters α (at $n = 5.25$, $\sigma = 1/2, m = 1 - 1/n$, $s_r = 0.17, s_s = 1$ and $K = 10^{-4}$ms^{-1})

is the *capillary diffusivity*. However, we will see further below that the formulation (10.14) is restricted to homogeneous porous media. With $C^{-1} = \partial\psi/\partial s$ we denote the *inverse moisture capacity* as a function of the saturation s. Analytic relations for $C^{-1}(s)$ are given in Appendix D. The porous-medium diffusivity $\boldsymbol{D}$ is generally a strongly nonlinear function of s (Fig. 10.3) and can vary drastically over space. Any solution method based on the $s-$form of the Richards' equation is restricted to unsaturated flow conditions because the saturation variable s is not unique for saturated regions, where the porous-medium diffusivity $\boldsymbol{D}$ goes to infinity and a retention relation $s(\psi)$ no longer exists. Moreover, it is important to note that the saturation s is basically a discontinuous variable because $s(\psi)$ is associated with the property of the porous medium which can abruptly vary in space, e.g., at material interfaces. Hence, the variable s is generally nonunique for heterogeneous porous media. That means that the application of the chain rule for $\nabla\psi$ in the form of (10.13) is mathematically not correct and the commonly derived $s-$form of the Richards' equation (10.14) is only valid for homogeneous porous media, unless specific material-gradient terms are added.[2] On the other hand, it has been shown, e.g., [249], that a

[2] The saturation relation $s(\psi)$ depends on the porous-medium properties, such as parameters α, n and m appearing in the van Genuchten relationship (D.4). In heterogeneous media the parameters can vary in space, i.e., $\alpha = \alpha(\boldsymbol{x})$, $n = n(\boldsymbol{x})$ and $m = m(\boldsymbol{x})$. Then, the chain rule applied to ∇s yields for a van Genuchten relationship

$$\nabla s = \frac{\partial s}{\partial \psi}\nabla\psi + \frac{\partial s}{\partial \alpha}\nabla\alpha + \frac{\partial s}{\partial n}\nabla n + \frac{\partial s}{\partial m}\nabla m$$

Table 10.2 Forms of Richards' equation written in terms of the variables ψ and s

Form	Equation	Primary variable	Modeling features
Mixed $\psi - s-$form	$s\, S_o \frac{\partial \psi}{\partial t} + \varepsilon \frac{\partial s}{\partial t}$ $-\nabla \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla \psi + (1+\chi)\boldsymbol{e})]$ $= Q_h + Q_{hw} + Q_{\mathrm{EOB}}$	ψ (or s)[a]	Most general Mass-conservative Unsaturated-saturated Heterogeneous porous media
$\psi-$form	$(s\, S_o + \varepsilon C) \frac{\partial \psi}{\partial t}$ $-\nabla \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla \psi + (1+\chi)\boldsymbol{e})]$ $= Q_h + Q_{hw} + Q_{\mathrm{EOB}}$	ψ	Standard, robust Poorly conservative Unsaturated-saturated Heterogeneous porous media
$s-$form	$(s\, S_o\, C^{-1} + \varepsilon) \frac{\partial s}{\partial t}$ $-\nabla \cdot [\boldsymbol{D} \cdot \nabla s + k_r \boldsymbol{K} f_\mu (1+\chi)\boldsymbol{e}]$ $= Q_h + Q_{hw} + Q_{\mathrm{EOB}}$	s	Only unsaturated, $s < 1$ Mass-conservative Efficient for dry conditions Homogeneous porous media

[a] In case of primary variable switching

$s-$based algorithm can result in significantly improved performances compared to $\psi-$based (or $h-$based) methods. This is due to the fact that the inherent parametric functions are less nonlinear when expressed in terms of s rather than ψ (or h), particularly when applied to relatively dry porous media.

Table 10.2 summarizes the three essential forms of the Richards' equation with their most important modeling features. It is obvious that only $\psi - s-$based and $\psi-$based forms are appropriate for computing simultaneous unsaturated-saturated flow in heterogeneous porous media. Equivalent formulations result when the hydraulic head h is used instead of the pressure head ψ.

and contrary to (10.14), the correct $s-$form of the Richards' equation reads for heterogeneous porous media:

$$\big(s\, S_o\, C^{-1} + \varepsilon\big) \frac{\partial s}{\partial t} - \nabla \cdot \big[\boldsymbol{D} \cdot (\nabla s - \tfrac{\partial s}{\partial \alpha} \nabla \alpha - \tfrac{\partial s}{\partial n} \nabla n - \tfrac{\partial s}{\partial m} \nabla m) + k_r \boldsymbol{K} f_\mu (1+\chi)\boldsymbol{e}\big] = Q_h + Q_{hw} + Q_{\mathrm{EOB}}$$

exemplified for a van Genuchten relationship. Similar expressions result for other empirical $s(\psi)-$relations, see Appendix D. The terms $\frac{\partial s}{\partial \alpha} \nabla \alpha$, $\frac{\partial s}{\partial n} \nabla n$ and $\frac{\partial s}{\partial m} \nabla m$ additionally appearing in the $s-$based form of the Richards' equation need a specific treatment in the numerical solution, e.g., [311]. More discussions are given by LaBolle and Clausnitzer [327].

10.4 Use of Transformation Methods

The objective of any transformation applied to the basic Richards' equation in its $\psi - s$–form (or equivalent $h - s$–form) is to find a formulation that will result in a more efficient and robust solution to (10.5). While the s–form represents a possibility of transformation, there are more useful transformation approaches [566] which can be utilized to significantly enhance the efficiency and accuracy of the numerical solution of the Richards' equation. Unfortunately, most of the transformation approaches necessitate restrictions in the formulation. Nevertheless, they can represent interesting alternatives to the standard solution strategies which can be inefficient and unreliable in solving the highly nonlinear flow problems in unsaturated porous media.

For specific needs an important transformation represents the *Kirchhoff integral transform* related to the relative permeability k_r written in the form [449]

$$F(\psi) = \int_{-\infty}^{\psi} k_r(\tau)d\tau \tag{10.16}$$

By using the exponential relationship (D.39) of Appendix D

$$k_r(\psi) = e^{\alpha\psi} \tag{10.17}$$

assuming an air-entry pressure head of $\psi_a = 0$, we obtain from (10.16)

$$F(\psi) = \int_{-\infty}^{\psi} e^{\alpha\tau} d\tau = \tfrac{1}{\alpha} e^{\alpha\tau}\big|_{-\infty}^{\psi} = \tfrac{1}{\alpha} e^{\alpha\psi} = \tfrac{1}{\alpha} k_r(\psi) \tag{10.18}$$

Now, if we assume that the sorptive parameter α is constant, i.e., the approach is restricted to homogeneous porous media, where the unsaturated parameter α is spatially invariable, we find

$$\nabla F = e^{\alpha\psi}\nabla F = k_r(\psi)\nabla F, \quad \frac{\partial F}{\partial t} = e^{\alpha\psi}\frac{\partial\psi}{\partial t}, \quad \frac{\partial F}{\partial t} = e^{\alpha\psi}C^{-1}\frac{\partial s}{\partial t} \tag{10.19}$$

and

$$k_r(\psi) = \alpha F \tag{10.20}$$

Then, the Richards' equation (10.5) takes the form

$$(s\,S_o + \varepsilon C)e^{-\alpha\psi}\frac{\partial F}{\partial t} - \nabla\cdot\big[\boldsymbol{K} f_\mu \cdot \underbrace{(k_r\nabla\psi)}_{\nabla F} + \boldsymbol{K} f_\mu \cdot \underbrace{k_r}_{\alpha F}(1+\chi)\boldsymbol{e})\big] = Q_h + Q_{hw} + Q_{\text{EOB}} \tag{10.21}$$

By using the capacity $C = \alpha(s_s - s_r)e^{\alpha\psi}$ for the exponential relationship (D.18) and additionally assume that the storage effects S_o are negligible, we can *linearize* the Richards' equation in the form

$$\varepsilon\alpha(s_s - s_r)\frac{\partial F}{\partial t} - \nabla \cdot (\boldsymbol{K} f_\mu \cdot \nabla F - \boldsymbol{v} F) = Q_h + Q_{hw} + Q_{\mathrm{EOB}} \tag{10.22}$$

with the capillary advection vector

$$\boldsymbol{v} = -\alpha \boldsymbol{K} f_\mu \cdot (1 + \chi)\boldsymbol{e} \tag{10.23}$$

for solving the transform function $0 < F < \frac{1}{\alpha}$. We note that (10.22) represents a divergence-form ADE which is linear in F, however, restricted to unsaturated conditions $s < 1$ and homogeneous (α−constant) porous media. Such a type of ADE can be solved with transformed BC's and IC[3] in a very efficiently and fast way by using standard techniques. After solving (10.22) for F in Ω for the given BC's and IC, the pressure head ψ and hydraulic head h can be simply obtained by backtransformation. Since $\alpha F = e^{\alpha\psi}$ and $\ln(\alpha F) = \alpha\psi$, we get

$$\begin{aligned} \psi &= \tfrac{1}{\alpha}\ln(\alpha F) \\ h &= \psi + z \end{aligned} \tag{10.24}$$

While s−transformed and Kirchhoff-integral-transformed formulations, (10.14) and (10.22), respectively, have shown powerful for special applications in unsaturated flow modeling, they are unfortunately not general enough for solving simultaneous unsaturated-saturated flow in heterogeneous porous media. To overcome these limitations we will introduce the *primary variable switching technique* (PVST) in Sect. 10.7 as a specific transformation strategy which can handle both

[3] BC's for the transformed ADE (10.22) can be equivalently found for (10.6) when written by the new F variable:

$$\begin{aligned} F &= \tfrac{1}{\alpha}e^{\alpha(h_D - z)} && \text{on} \quad \Gamma_D \times t[t_0, \infty) \\ -(\boldsymbol{K} f_\mu \cdot \nabla F - \boldsymbol{v} F) \cdot \boldsymbol{n} &= q_F && \text{on} \quad \Gamma_N \times t[t_0, \infty) \\ -[\boldsymbol{K} f_\mu \cdot (1 + \chi)\boldsymbol{e}] \cdot \boldsymbol{n} &= \tfrac{1}{\alpha}\boldsymbol{v} \cdot \boldsymbol{n} = q_F^{\nabla} && \text{on} \quad \Gamma_N^{\nabla} \times t[t_0, \infty) \\ -(\boldsymbol{K} f_\mu \cdot \nabla F - \boldsymbol{v} F) \cdot \boldsymbol{n} &= -\Phi_h[h_C - z - \tfrac{1}{\alpha}\ln(\alpha F)] && \text{on} \quad \Gamma_C \times t[t_0, \infty) \\ Q_{hw} &= -\textstyle\sum_w Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty) \end{aligned}$$

additionally, the seepage face BC for (10.7) as

$$F = \tfrac{1}{\alpha} \quad \text{at} \quad Q_{n_h} > 0 \quad \text{on} \quad \Gamma_S \times t[t_0, \infty)$$

and the IC (10.8) in the form

$$F(\boldsymbol{x}, t_0) = \tfrac{1}{\alpha}e^{\alpha[h_0(\boldsymbol{x}) - z]} \quad \text{in} \quad \bar{\Omega}$$

We note that the Cauchy-type BC on Γ_C introduces a nonlinear expression in F.

unsaturated and saturated flow conditions very efficiently without restriction to homogeneous media.

10.5 Finite Element Formulation of the Mixed $h - s$—Based Form of Richards' Equation

Based on the principles of FEM thoroughly described in Chap. 8 we apply now the GFEM to the basic Richards' equation in its mixed form stated above in Sect. 10.2 associated with the corresponding BC's and IC's for solving simultaneous unsaturated-saturated flow in heterogeneous porous media. For convenience we focus firstly on the finite element formulation of the mixed $h - s$—form of the Richards' equation for multidimensional problems (i.e., 3D, 2D and axisymmetric). The formulation for the alternative ψ—s—form of the Richards' equation will appear rather similar to the h—s—form. The standard h—form (and the equivalent ψ—form) of the Richards' equation can be easily deduced from the given developments for the $h - s$—form. However, the specific s—form and the Kirchhoff-integral-transformed formulation of the Richards' equation are not followed here due to the inherent restrictions as discussed above in Sects. 10.3 and 10.4, respectively. Instead, the PVST is preferred which will be thoroughly discussed in Sect. 10.7.

10.5.1 Weak Form

According to Sect. 8.5 the weak form for the h—s—based Richards' equation (10.5) appears as a special case of the ADE weak statement deduced from the expression (8.53). We find

$$\begin{aligned}\int_\Omega w s S_o \frac{\partial h}{\partial t} d\Omega + \int_\Omega w \varepsilon \frac{\partial s}{\partial t} d\Omega + \\ \int_\Omega \nabla w \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega - \int_\Omega w (Q_h + Q_{hw} + Q_{\text{EOB}}) d\Omega - \\ \int_\Gamma w [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] \cdot \boldsymbol{n}\, d\Gamma = 0, \quad \forall w \in H^1(\Omega) \qquad (10.25)\end{aligned}$$

where w is a suitable weighting function. Separating the boundary integral of (10.25) into the five segments $\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_N^\nabla \cup \Gamma_C$ imposed by the Dirichlet, Neumann, gradient and Cauchy-type BC's, respectively, we invoke the BC's and SPC of (10.6) and BC of (10.7) to obtain

$$
\begin{aligned}
&\int_{\Omega} w s S_o \frac{\partial h}{\partial t} d\Omega + \int_{\Omega} w \varepsilon \frac{\partial s}{\partial t} d\Omega + \\
&\int_{\Omega} \nabla w \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega - \int_{\Omega} w (Q_h + Q_{\mathrm{EOB}}) d\Omega + \\
&\sum_w w(\boldsymbol{x}_w) Q_w(t) + \int_{\Gamma_N} w q_h d\Gamma + \int_{\Gamma_N^{\nabla}} w k_r q_h^{\nabla} d\Gamma - \\
&\int_{\Gamma_C} w \Phi_h (h_C - h) d\Gamma = 0, \qquad \forall w \in H_0^1(\Omega)
\end{aligned}
\tag{10.26}
$$

10.5.2 GFEM and Resulting Nonlinear Matrix System

The weak statement of the $h - s-$based form of the Richards' equation (10.26) involves the two unknown variables h and s. In using the FEM these variables are replaced by a *continuous approximation* that assumes the separability of space and time (see Sect. 8.4). Thus

$$
\left.
\begin{aligned}
h(\boldsymbol{x}, t) &\approx \textstyle\sum_j N_j(\boldsymbol{x}) h_j(t) \\
s(\boldsymbol{x}, t) &\approx \textstyle\sum_j N_j(\boldsymbol{x}) s_j(t)
\end{aligned}
\right\} \quad j = 1, \ldots, N_{\mathrm{P}}
\tag{10.27}
$$

where j designates global nodal indices. It is important to emphasize that also the saturation variable s, although basically discontinuous in heterogeneous media, is approximated in a continuous manner. Now, using the Galerkin method with the weighting function

$$
w \rightarrow w_i = N_i, \quad i = 1, \ldots, N_{\mathrm{P}}
\tag{10.28}
$$

we find the following Galerkin-based finite element formulation of (10.26), viz.,

$$
\begin{aligned}
&\sum_e \int_{\Omega^e} N_i s^e S_o^e \frac{\partial}{\partial t} (\sum_j N_j h_j) d\Omega^e + \sum_e \int_{\Omega^e} N_i \varepsilon^e \frac{\partial}{\partial t} (\sum_j N_j s_j) d\Omega^e + \\
&\sum_e \int_{\Omega^e} \nabla N_i \cdot \big[k_r^e \boldsymbol{K}^e f_\mu^e \cdot \nabla (\sum_j N_j h_j)\big] d\Omega^e + \sum_e \int_{\Omega^e} \nabla N_i \cdot \big(k_r^e \boldsymbol{K}^e f_\mu^e \cdot \chi^e \boldsymbol{e}\big) d\Omega^e - \\
&\sum_e \int_{\Omega^e} N_i (Q_h^e + Q_{\mathrm{EOB}}^e) d\Omega^e + Q_w(t)\big|_i + \\
&\sum_e \int_{\Gamma_N^e} N_i q_h^e d\Gamma^e + \sum_e \int_{\Gamma_N^{\nabla e}} N_i k_r^e q_h^{\nabla^e} d\Gamma^e - \sum_e \int_{\Gamma_C^e} N_i \Phi_h^e \big[h_C^e - (\sum_j N_j h_j)\big] d\Gamma^e = 0 \\
&1 \le i, j \le N_{\mathrm{P}}
\end{aligned}
\tag{10.29}
$$

The assembly process leads the nonlinear global matrix system of N_{P} equations

$$\boldsymbol{O}(\boldsymbol{s}) \cdot \dot{\boldsymbol{h}} + \boldsymbol{B} \cdot \dot{\boldsymbol{s}} + \boldsymbol{D}(\boldsymbol{s}) \cdot \boldsymbol{h} - \boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0} \tag{10.30}$$

showing the nonlinearities in parentheses, where

$$\boldsymbol{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_{N_{\mathrm{P}}} \end{pmatrix}, \quad \boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{N_{\mathrm{P}}} \end{pmatrix}, \quad \dot{\boldsymbol{h}} = \begin{pmatrix} \frac{dh_1}{dt} \\ \frac{dh_2}{dt} \\ \vdots \\ \frac{dh_{N_{\mathrm{P}}}}{dt} \end{pmatrix}, \quad \dot{\boldsymbol{s}} = \begin{pmatrix} \frac{ds_1}{dt} \\ \frac{ds_2}{dt} \\ \vdots \\ \frac{ds_{N_{\mathrm{P}}}}{dt} \end{pmatrix} \tag{10.31}$$

and the matrices and RHS vector

$$\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \delta_{ij} \int_{\Omega^e} s^e S_o^e \, N_i \, d\Omega^e \\
\boldsymbol{B} = B_{ij} &= \sum_e \delta_{ij} \int_{\Omega^e} \varepsilon^e \, N_i \, d\Omega^e \\
\boldsymbol{D} = D_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot \big(k_r^e(s^e) \boldsymbol{K}^e f_\mu^e \cdot \nabla N_j \big) d\Omega^e + \int_{\Gamma_C^e} \Phi_h^e N_i N_j \, d\Gamma^e \Big) \\
\boldsymbol{F} = F_i \; &= \sum_e \Big(\int_{\Omega^e} N_i (Q_h^e + Q_{\mathrm{EOB}}^e) d\Omega^e - \\
&\quad \int_{\Omega^e} \nabla N_i \cdot \big(k_r^e(s^e) \boldsymbol{K}^e f_\mu^e \cdot \chi^e \boldsymbol{e} \big) d\Omega^e + \int_{\Gamma_C^e} N_i \Phi_h^e h_C^e \, d\Gamma^e - \\
&\quad \int_{\Gamma_N^e} N_i q_h^e d\Gamma^e - \int_{\Gamma_N^{\nabla e}} N_i k_r^e(s^e) q_h^{\nabla e} \, d\Gamma^e \Big) - Q_w(t)\big|_i
\end{aligned} \tag{10.32}$$

for $(i, j = 1, \ldots, N_{\mathrm{P}})$ and $(e = 1, \ldots, N_{\mathrm{E}})$, in which the following interpolations for the saturation and relative permeability over the element e are employed

$$s^e = \sum_J N_J^e s_J^e, \quad k_r^e(s^e) = \sum_J N_J^e k_r^e(s_J^e) \tag{10.33}$$

where J runs over local node numbers. In (10.33) the saturation s_J^e at local node J of element e can be evaluated from the retention relations $s_J^e = f(\psi_J^e)$ for given pressure heads $\psi_J^e = h_J^e - z_J$ which are determined from the $h-$solution. Note that the matrices $\boldsymbol{O}$ and $\boldsymbol{B}$ connected with time derivatives are mass-lumped (cf. Sect. 8.13.2), where δ_{ij} is the Kronecker symbol applied to global indices. This is virtually mandatory for unsaturated problems to ensure smooth and non-oscillatory solutions, e.g., [72, 295]. It is important to note that $\boldsymbol{D}$ is *symmetric* and accordingly the complete matrix system (10.30) is symmetric. On the other hand, the system of equations (10.30) is highly nonlinear due to the functional dependence of the constitutive relationships for the saturation and the relative permeability. The integrals appearing in (10.32) are performed on element level in the local coordinates (see Sect. 8.12). Analytical evaluations of partial integral terms of (10.32) can be deduced from developments done in Appendix H for selected element types. The differential elements $d\Omega^e$ and $d\Gamma^e$ differ for 3D, 2D and axisymmetric problems as given by (8.122)–(8.124), respectively. The tensor of the saturated hydraulic conductivity $\boldsymbol{K}^e$ of element e may be anisotropic in formulations introduced in Chap. 7.

10.5.3 Time Integration and Celia et al.'s Approximation Method of Picard Iteration

For solving the nonlinear matrix system (10.30) in time t with the associated IC an appropriate time marching recurrence scheme combined with an iteration strategy has to be applied such as introduced in Sects. 8.13 and 8.18, respectively. For the unsaturated-saturated flow problems based on the $h - s-$form of the Richards' equation we find the temporally discretized formulation of (10.30) as

$$\boldsymbol{O}(\boldsymbol{s}_{n+1}) \cdot \Big(\frac{\boldsymbol{h}_{n+1} - \boldsymbol{h}_n}{\theta \Delta t_n} - (\tfrac{1}{\theta} - 1)\dot{\boldsymbol{h}}_n \Big) + \boldsymbol{B} \cdot \Big(\frac{\boldsymbol{s}_{n+1} - \boldsymbol{s}_n}{\theta \Delta t_n} - (\tfrac{1}{\theta} - 1)\dot{\boldsymbol{s}}_n \Big) + \\ \boldsymbol{D}(\boldsymbol{s}_{n+1}) \cdot \boldsymbol{h}_{n+1} - \boldsymbol{F}(\boldsymbol{s}_{n+1}) = \boldsymbol{0} \tag{10.34}$$

where $\theta \in (\frac{1}{2}, 1)$ for the TR and BE scheme, respectively. Commonly, fully implicit time integration with $\theta = 1$ is preferred due to its robustness. The Picard iteration method (Sect. 8.18.1) can be the first choice for the iterative solution of (10.34). It results in

$$\boldsymbol{O}(\boldsymbol{s}^{\tau}_{n+1}) \cdot \Big(\frac{\boldsymbol{h}^{\tau+1}_{n+1} - \boldsymbol{h}_n}{\theta \Delta t_n} - (\tfrac{1}{\theta} - 1)\dot{\boldsymbol{h}}_n \Big) + \boldsymbol{B} \cdot \Big(\frac{\boldsymbol{s}^{\tau+1}_{n+1} - \boldsymbol{s}_n}{\theta \Delta t_n} - (\tfrac{1}{\theta} - 1)\dot{\boldsymbol{s}}_n \Big) + \\ \boldsymbol{D}(\boldsymbol{s}^{\tau}_{n+1}) \cdot \boldsymbol{h}^{\tau+1}_{n+1} - \boldsymbol{F}(\boldsymbol{s}^{\tau}_{n+1}) = \boldsymbol{0} \tag{10.35}$$

where $\tau = 0, 1, \ldots$ is the iteration counter. We note that the Picard method has the advantage to preserve the symmetry of the resulting discrete system of flow equations. Since the matrix system still involves the two variables $\boldsymbol{h}$ and $\boldsymbol{s}$, it is necessary to replace one variable by the other one. But, this must be done suitably to avoid mass balance errors in the approximation. A precise mass-conservative method has been proposed by Celia et al. [72], in which $\boldsymbol{s}^{\tau+1}_{n+1}$ is expanded in a truncated Taylor series with respect to $\boldsymbol{h}$ about the expansion point $\boldsymbol{h}^{\tau}_{n+1}$ in the following form:

$$\boldsymbol{s}^{\tau+1}_{n+1} = \boldsymbol{s}^{\tau}_{n+1} + \frac{\partial \boldsymbol{s}^{\tau}_{n+1}}{\partial \boldsymbol{h}^{\tau}_{n+1}} \cdot \big(\boldsymbol{h}^{\tau+1}_{n+1} - \boldsymbol{h}^{\tau}_{n+1} \big) + \text{HOT} \tag{10.36}$$

After neglecting all terms higher than linear (HOT ≈ 0), it results

$$\frac{\boldsymbol{s}_{n+1}^{\tau+1}-\boldsymbol{s}_n}{\theta\Delta t_n} = \boldsymbol{C}_{n+1}^{\tau}\cdot\frac{\boldsymbol{h}_{n+1}^{\tau+1}-\boldsymbol{h}_{n+1}^{\tau}}{\theta\Delta t_n}+\frac{\boldsymbol{s}_{n+1}^{\tau}-\boldsymbol{s}_n}{\theta\Delta t_n} \tag{10.37}$$

where

$$\boldsymbol{C}_{n+1}^{\tau} = \frac{\partial\boldsymbol{s}_{n+1}^{\tau}}{\partial\boldsymbol{h}_{n+1}^{\tau}} = \frac{\partial\boldsymbol{s}_{n+1}^{\tau}}{\partial\boldsymbol{\psi}_{n+1}^{\tau}} \tag{10.38}$$

is the moisture capacity matrix which can be evaluated at each discrete pressure head $\boldsymbol{\psi} = \boldsymbol{h}-\boldsymbol{z}$ for given time plane $n+1$ and iteration τ of analytical or numerical (chord slope) C—relationships as described in Appendix D or in Sect. J.3 of Appendix J, respectively. Substituting (10.37) into (10.35) we obtain the Celia et al.'s mass-conservative Picard-type iteration method for the mixed $h-s$—based form of the Richards' equation as

$$\begin{aligned}\left(\frac{\boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau})}{\theta\Delta t_n}+\frac{\boldsymbol{B}}{\theta\Delta t_n}\cdot\boldsymbol{C}_{n+1}^{\tau}+\boldsymbol{D}(\boldsymbol{s}_{n+1}^{\tau})\right)\cdot\boldsymbol{h}_{n+1}^{\tau+1} = \\ \boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau})\cdot\left(\frac{\boldsymbol{h}_n}{\theta\Delta t_n}+(\tfrac{1}{\theta}-1)\dot{\boldsymbol{h}}_n\right)+\left(\frac{\boldsymbol{B}}{\theta\Delta t_n}\cdot\boldsymbol{C}_{n+1}^{\tau}\right)\cdot\boldsymbol{h}_{n+1}^{\tau}+ \\ \boldsymbol{B}\cdot\left((\tfrac{1}{\theta}-1)\dot{\boldsymbol{s}}_n-\frac{\boldsymbol{s}_{n+1}^{\tau}-\boldsymbol{s}_n}{\theta\Delta t_n}\right)+\boldsymbol{F}(\boldsymbol{s}_{n+1}^{\tau})\end{aligned} \tag{10.39}$$

to solve the unknown vector of hydraulic head $\boldsymbol{h}_{n+1}^{\tau+1}$ as the primary variable at new iteration $\tau+1$ and new time plane $n+1$. Within the iteration τ and at the given time stage $n+1$, the saturation vector $\boldsymbol{s}_{n+1}^{\tau}$ is taken from the previous iterate $\boldsymbol{h}_{n+1}^{\tau}$ by evaluating the retention relationship $f(\boldsymbol{\psi}_{n+1}^{\tau})$ of Appendix D with $\boldsymbol{\psi}_{n+1}^{\tau} = \boldsymbol{h}_{n+1}^{\tau}-\boldsymbol{z}$ for each node.

The Picard iteration in Celia et al.'s linearization (10.39) is usually terminated via a deviatory error criterion, such as

$$\frac{\|\boldsymbol{h}_{n+1}^{\tau+1}-\boldsymbol{h}_{n+1}^{\tau}\|}{\|\boldsymbol{h}_{n+1}^{\tau+1}\|}\le\epsilon \tag{10.40}$$

where ϵ is a defined error tolerance. Under certain conditions the convergence test (10.40) has shown too rough when changes in $\boldsymbol{h}$ are small and smoothly behaved while changes in the saturation $\boldsymbol{s}$ and/or pressure head $\boldsymbol{\psi}$ remain significant. A remedy could be the exacerbated convergence criterion

$$\max\left(\frac{\|\boldsymbol{h}_{n+1}^{\tau+1}-\boldsymbol{h}_{n+1}^{\tau}\|}{\|\boldsymbol{h}_{n+1}^{\tau+1}\|},\frac{\|\boldsymbol{\psi}_{n+1}^{\tau+1}-\boldsymbol{\psi}_{n+1}^{\tau}\|}{\|\boldsymbol{\psi}_{n+1}^{\tau+1}\|},\frac{\|\boldsymbol{s}_{n+1}^{\tau+1}-\boldsymbol{s}_{n+1}^{\tau}\|}{\|\boldsymbol{s}_{n+1}^{\tau+1}\|}\right)\le\epsilon \tag{10.41}$$

where the pressure head $\psi = \boldsymbol{h} - \boldsymbol{z}$ and the saturation $\boldsymbol{s} = f(\boldsymbol{\psi})$ are evaluated from the $\boldsymbol{h}-$solution.

In using FEFLOW's predictor-corrector time integration with automatic error-controlled time stepping a one-step Picard method (cf. Sect. 8.18.4) for transient unsaturated-saturated flow problems have shown powerful [141], in which the predictor solutions $\boldsymbol{h}_{n+1}^p$, $\boldsymbol{\psi}_{n+1}^p$ and $\boldsymbol{s}_{n+1}^p$ are used to linearize (10.39) in the form

$$\begin{aligned}\Big(\frac{\boldsymbol{O}(\boldsymbol{s}_{n+1}^p)}{\theta\Delta t_n} + \frac{\boldsymbol{B}}{\theta\Delta t_n}\cdot\boldsymbol{C}_{n+1}^p + \boldsymbol{D}(\boldsymbol{s}_{n+1}^p)\Big)\cdot\boldsymbol{h}_{n+1} = \\ \boldsymbol{O}(\boldsymbol{s}_{n+1}^p)\cdot\Big(\frac{\boldsymbol{h}_n}{\theta\Delta t_n} + (\tfrac{1}{\theta}-1)\dot{\boldsymbol{h}}_n\Big) + \Big(\frac{\boldsymbol{B}}{\theta\Delta t_n}\cdot\boldsymbol{C}_{n+1}^p\Big)\cdot\boldsymbol{h}_{n+1}^p + \\ \boldsymbol{B}\cdot\Big((\tfrac{1}{\theta}-1)\dot{\boldsymbol{s}}_n - \frac{\boldsymbol{s}_{n+1}^p - \boldsymbol{s}_n}{\theta\Delta t_n}\Big) + \boldsymbol{F}(\boldsymbol{s}_{n+1}^p)\end{aligned} \tag{10.42}$$

in solving $\boldsymbol{h}_{n+1}$ at the new time plane $n+1$, where

$$\begin{aligned}\boldsymbol{h}_{n+1}^p &= \begin{cases}\boldsymbol{h}_n + \Delta t_n\dot{\boldsymbol{h}}_n\\ \boldsymbol{h}_n + \frac{\Delta t_n}{2}\big[(2+\frac{\Delta t_n}{\Delta t_{n-1}})\dot{\boldsymbol{h}}_n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\boldsymbol{h}}_{n-1}\big]\end{cases}\\ \boldsymbol{\psi}_{n+1}^p &= \begin{cases}\boldsymbol{\psi}_n + \Delta t_n\dot{\boldsymbol{\psi}}_n\\ \boldsymbol{\psi}_n + \frac{\Delta t_n}{2}\big[(2+\frac{\Delta t_n}{\Delta t_{n-1}})\dot{\boldsymbol{\psi}}_n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\boldsymbol{\psi}}_{n-1}\big]\end{cases} \quad \text{with} \quad \boldsymbol{\psi}_n = \boldsymbol{h}_n - \boldsymbol{z}\\ \boldsymbol{s}_{n+1}^p &= \begin{cases}\boldsymbol{s}_n + \Delta t_n\dot{\boldsymbol{s}}_n\\ \boldsymbol{s}_n + \frac{\Delta t_n}{2}\big[(2+\frac{\Delta t_n}{\Delta t_{n-1}})\dot{\boldsymbol{s}}_n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\boldsymbol{s}}_{n-1}\big]\end{cases} \quad \text{with} \quad \boldsymbol{s}_n = f(\boldsymbol{\psi}_n)\\ \boldsymbol{C}_{n+1}^p &= \frac{\partial\boldsymbol{s}_{n+1}^p}{\partial\boldsymbol{\psi}_{n+1}^p}\end{aligned} \tag{10.43}$$

by using the suited FE and AB predictors (cf. Sect. 8.13.5 and Table 8.7), respectively, in which the acceleration vectors

$$\begin{aligned}\dot{\boldsymbol{h}}_n &= \begin{cases}\frac{\boldsymbol{h}_n-\boldsymbol{h}_{n-1}}{\Delta t_{n-1}}\\ \big(2-\frac{\Delta t_{n-2}}{\Delta t_{n-1}+\Delta t_{n-2}}\big)\big(\frac{\boldsymbol{h}_n-\boldsymbol{h}_{n-1}}{\Delta t_{n-1}}\big) - \big(\frac{\Delta t_{n-1}}{\Delta t_{n-1}+\Delta t_{n-2}}\big)\big(\frac{\boldsymbol{h}_{n-1}-\boldsymbol{h}_{n-2}}{\Delta t_{n-2}}\big)\end{cases}\\ \dot{\boldsymbol{\psi}}_n &= \begin{cases}\frac{\boldsymbol{\psi}_n-\boldsymbol{\psi}_{n-1}}{\Delta t_{n-1}}\\ \big(2-\frac{\Delta t_{n-2}}{\Delta t_{n-1}+\Delta t_{n-2}}\big)\big(\frac{\boldsymbol{\psi}_n-\boldsymbol{\psi}_{n-1}}{\Delta t_{n-1}}\big) - \big(\frac{\Delta t_{n-1}}{\Delta t_{n-1}+\Delta t_{n-2}}\big)\big(\frac{\boldsymbol{\psi}_{n-1}-\boldsymbol{\psi}_{n-2}}{\Delta t_{n-2}}\big)\end{cases}\\ \dot{\boldsymbol{s}}_n &= \begin{cases}\frac{\boldsymbol{s}_n-\boldsymbol{s}_{n-1}}{\Delta t_{n-1}}\\ \big(2-\frac{\Delta t_{n-2}}{\Delta t_{n-1}+\Delta t_{n-2}}\big)\big(\frac{\boldsymbol{s}_n-\boldsymbol{s}_{n-1}}{\Delta t_{n-1}}\big) - \big(\frac{\Delta t_{n-1}}{\Delta t_{n-1}+\Delta t_{n-2}}\big)\big(\frac{\boldsymbol{s}_{n-1}-\boldsymbol{s}_{n-2}}{\Delta t_{n-2}}\big)\end{cases}\end{aligned} \tag{10.44}$$

have to be recorded during the time stepping procedure for the FE and AB predictors, respectively. The following deviatory error estimates (cf. Table 8.7)

$$\begin{aligned} \boldsymbol{d}^h_{n+1} &= \begin{cases} \frac{1}{2}(\boldsymbol{h}_{n+1} - \boldsymbol{h}^p_{n+1}) \\ \frac{1}{3}(\boldsymbol{h}_{n+1} - \boldsymbol{h}^p_{n+1})/(1 + \frac{\Delta t_{n-1}}{\Delta t_n}) \end{cases} \\ \boldsymbol{d}^\psi_{n+1} &= \begin{cases} \frac{1}{2}(\boldsymbol{\psi}_{n+1} - \boldsymbol{\psi}^p_{n+1}) \\ \frac{1}{3}(\boldsymbol{\psi}_{n+1} - \boldsymbol{\psi}^p_{n+1})/(1 + \frac{\Delta t_{n-1}}{\Delta t_n}) \end{cases} \\ \boldsymbol{d}^s_{n+1} &= \begin{cases} \frac{1}{2}(\boldsymbol{s}_{n+1} - \boldsymbol{s}^p_{n+1}) \\ \frac{1}{3}(\boldsymbol{s}_{n+1} - \boldsymbol{s}^p_{n+1})/(1 + \frac{\Delta t_{n-1}}{\Delta t_n}) \end{cases} \end{aligned} \tag{10.45}$$

are used in the time step control of the FE/BE and AB/TR predictor-corrector schemes, respectively, according to

$$\Delta t_{n+1} = \begin{cases} \Delta t_n \left(\dfrac{\epsilon}{\|\boldsymbol{d}^h_{n+1}\|} \right)^{1/\lambda} & \text{or optionally} \\ \Delta t_n \left(\dfrac{\epsilon}{\max(\|\boldsymbol{d}^h_{n+1}\|, \|\boldsymbol{d}^\psi_{n+1}\|, \|\boldsymbol{d}^s_{n+1}\|)} \right)^{1/\lambda} \end{cases} \tag{10.46}$$

where $\lambda = 2$ for the FE/BE scheme, $\lambda = 3$ for the AB/TR scheme and ϵ is the pre-set error tolerance measure.

We note that Celia et al.'s approximation method is only one option in FEFLOW. A drawback is its restriction to a Picard-type iteration which has only a linear convergence rate. In the subsequent Sect. 10.7 we will introduce a generalization in form of PVST established with the full Newton iteration method of quadratic convergence, where we can show that Celia et al.'s linearization deduces from PVST as a special case. Furthermore, the deviatory error is often insufficient for a convergence control and an additional test of the mass balance error via a direct control of the discrete residuals seems appropriate in many cases. Indeed, PVST incorporates a family of methods. On PVST's basis we will develop mass-conservative modeling options encompassing the full Newton iteration method and the Picard iteration method with and without additional residual control as well as both 1st-order accurate fully implicit and 2nd-order accurate semi-implicit time integration methods for solving the mixed $h - s-$based (and the equivalent $\psi - s-$based) form of the governing Richards' equation.

10.6 Finite Element Formulation of the Standard $h-$Based Form of Richards' Equation

The standard formulation of the Richards' equation in form of (10.12) does not guarantee mass conservation in its discrete approximation due to the replacement of the storage term $\partial s/\partial t$ by the expansion $C\,\partial h/\partial t$, e.g., [72, 141]. Nevertheless, the standard formulation should not be generally rejected. In fact, it provides a high robustness and achieves reasonably accurate solutions for moderate saturation behaviors, in particular to compute seepage problems in phreatic aquifers in which the location of the free surface (as the zero-pressure interface) is of specific concern. Furthermore, it is always useful when major interest is in steady-state solution (whenever exists in unsaturated flow).

10.6.1 Spatial Discretization and Resulting Nonlinear Matrix System

The weak form for (10.12) with the involved BC's and SPC is equivalent to (10.26), except for the storage term. We find for the standard $h-$based form of the Richards' equation the following weak statement:

$$\begin{aligned}\int_{\Omega} w(sS_o + \varepsilon C)\frac{\partial h}{\partial t} d\Omega &+ \\ \int_{\Omega} \nabla w \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega - \int_{\Omega} w(Q_h + Q_{\text{EOB}}) d\Omega &+ \\ \sum_w w(\boldsymbol{x}_w) Q_w(t) + \int_{\Gamma_N} w q_h d\Gamma + \int_{\Gamma_N^{\nabla}} w k_r q_h^{\nabla} d\Gamma &- \\ \int_{\Gamma_C} w \Phi_h (h_C - h) d\Gamma = 0, \quad \forall w \in H_0^1(\Omega) &\end{aligned} \tag{10.47}$$

The only primary variable is h, which is approximated in the context of FEM:

$$h(\boldsymbol{x}, t) \approx \sum_j N_j(\boldsymbol{x})\, h_j(t), \quad j = 1, \ldots, N_{\text{P}} \tag{10.48}$$

Applying the Galerkin method with ($w \rightarrow w_i = N_i,\ i = 1, \ldots, N_{\text{P}}$) the weak statement (10.47) yields

$$\begin{aligned}\sum_e \int_{\Omega^e} N_i (s^e S_o^e + \varepsilon^e C^e) \frac{\partial}{\partial t} (\sum_j N_j h_j) d\Omega^e &+ \\ \sum_e \int_{\Omega^e} \nabla N_i \cdot \big[k_r^e \boldsymbol{K}^e f_\mu^e \cdot \nabla(\sum_j N_j h_j)\big] d\Omega^e + \sum_e \int_{\Omega^e} \nabla N_i \cdot \big(k_r^e \boldsymbol{K}^e f_\mu^e \cdot \chi^e \boldsymbol{e}\big) d\Omega^e &- \\ \sum_e \int_{\Omega^e} N_i (Q_h^e + Q_{\text{EOB}}^e) d\Omega^e + Q_w(t)\big|_i &+ \\ \sum_e \int_{\Gamma_N^e} N_i q_h^e d\Gamma^e + \sum_e \int_{\Gamma_N^{\nabla^e}} N_i k_r^e q_h^{\nabla^e} d\Gamma^e - \sum_e \int_{\Gamma_C^e} N_i \Phi_h^e \big[h_C^e - (\sum_j N_j h_j)\big] d\Gamma^e &= 0 \\ 1 \leq i, j \leq N_{\text{P}} &\end{aligned} \tag{10.49}$$

It leads to the following global symmetric matrix system of N_{P} equations

$$\boldsymbol{O}^{\dagger}(\boldsymbol{s}) \cdot \dot{\boldsymbol{h}} + \boldsymbol{D}(\boldsymbol{s}) \cdot \boldsymbol{h} - \boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0} \tag{10.50}$$

where the nonlinearities are shown in parentheses. The unknown vectors $\boldsymbol{h}$ and $\dot{\boldsymbol{h}}$ are equivalent to (10.31). The matrix $\boldsymbol{D}$ and the RHS vector $\boldsymbol{F}$ are given in (10.32). The only differences to the $h - s-$formulation of (10.30) are in the absence of the $\boldsymbol{B}-$storage matrix and in the modified form of the storage matrix, which reads now:

$$\boldsymbol{O}^{\dagger} = O_{ij}^{\dagger} = \sum_e \delta_{ij} \int_{\Omega^e} (s^e S_o^e + \varepsilon^e C^e)\, N_i \, d\Omega^e \tag{10.51}$$

written again in a mass-lumped formulation. In addition to (10.33) the interpolation of the moisture capacity over the element e is employed as

$$C^e(s^e) = \sum_J N_J^e C^e(s_J^e) \tag{10.52}$$

where the saturation s_J^e at local node J of element e is evaluated from the retention relations $s_J^e = f(\psi_J^e)$ for given pressure heads $\psi_J^e = h_J^e - z_J$ which are determined from the $h-$solution. The treatment of the integrals in $\boldsymbol{O}^{\dagger}$, $\boldsymbol{D}$ and $\boldsymbol{F}$ have been already discussed in the preceding Sect. 10.5.2.

The moisture capacity $C(s)$ can be evaluated both analytically and numerically. Analytic relations are summarized in Appendix D. Numerical evaluation of $C(s)$ can be performed by suited using chord slope approximations as discussed in Sect. J.3 of Appendix J. They are often preferred in the present $h-$form (and the equivalent $\psi-$form) of Richards' equation to improve global mass conservativity [435]. However, the numerical differentiation must be prevented if the hydraulic head difference falls below a specific range and a proper treatment of the derivative term is then required (for instance, resorting to an analytic evaluation). Accordingly, chord slope approximation does not appear as a general and sufficiently robust technique. It shall fail under drastic parameters and IC's [408]. Note that these difficulties are avoided when using mixed-form schemes in form of Celia et al.'s linearization (Sect. 10.5) and PVST (Sect. 10.7) which possess much better properties with respect to accurate mass conservative solutions.

10.6.2 Time Integration and Iteration Methods

The nonlinear matrix system (10.50) has to be solved in time t with the associated IC (10.8) via suitable single-step semi-implicit or fully implicit time marching recurrence schemes as described in Sect. 8.13. The GLS predictor-corrector time stepping combined with an automatic error-controlled time step selection strategy (cf. Sect. 8.13.5 and Table 8.7) is the preferred method

$$\Big(\frac{\boldsymbol{O}^{\dagger}(\boldsymbol{s}_{n+1})}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}_{n+1})\Big)\cdot\boldsymbol{h}_{n+1} = \boldsymbol{O}^{\dagger}(\boldsymbol{s}_{n+1})\cdot\Big(\frac{\boldsymbol{h}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{h}}_n\Big) + \boldsymbol{F}_{n+1}(\boldsymbol{s}_{n+1}) \qquad (10.53)$$

where $\theta \in (\frac{1}{2}, 1)$ for the TR and BE scheme, respectively. Alternatively, for user-defined (fixed) time step sizes Δt_n the $\theta-$method (Sect. 8.13.4) is useful

$$\Big(\frac{\boldsymbol{O}^{\dagger}(\boldsymbol{s}_{n+1})}{\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}_{n+1})\theta\Big)\cdot\boldsymbol{h}_{n+1} = \Big(\frac{\boldsymbol{O}^{\dagger}(\boldsymbol{s}_{n+1})}{\Delta t_n} - \boldsymbol{D}(\boldsymbol{s}_{n+1})(1-\theta)\Big)\cdot\boldsymbol{h}_n + \big(\boldsymbol{F}_{n+1}(\boldsymbol{s}_{n+1})\theta + \boldsymbol{F}_n(\boldsymbol{s}_n)(1-\theta)\big) \qquad (10.54)$$

where $\theta \in (\frac{1}{2}, \frac{2}{3}, 1)$ for the Crank-Nicolson, the Galerkin-in-time and the fully implicit scheme, respectively.

Iteration methods are required to solve (10.53) or (10.54) for $\boldsymbol{h}_{n+1}$. Most common is the Picard iteration method (cf. Sect. 8.18.4), which is computationally inexpensive, robust and preserves symmetry of the discrete system of flow equations, however, at the expense of only a linear convergence rate. For example, the Picard iteration method reads for the GLS predictor-corrector time integration in the general form[4] as

$$\Big(\frac{\boldsymbol{O}^{\dagger}(\boldsymbol{s}^{\tau}_{n+1})}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}^{\tau}_{n+1})\Big)\cdot\boldsymbol{h}^{\tau+1}_{n+1} = \boldsymbol{O}^{\dagger}(\boldsymbol{s}^{\tau}_{n+1})\cdot\Big(\frac{\boldsymbol{h}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{h}}_n\Big) + \boldsymbol{F}_{n+1}(\boldsymbol{s}^{\tau}_{n+1}) \qquad (10.55)$$

where $\tau = 0, 1, \ldots$ is the iteration counter, and in the one-step Picard form as

$$\Big(\frac{\boldsymbol{O}^{\dagger}(\boldsymbol{s}^{p}_{n+1})}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}^{p}_{n+1})\Big)\cdot\boldsymbol{h}_{n+1} = \boldsymbol{O}^{\dagger}(\boldsymbol{s}^{p}_{n+1})\cdot\Big(\frac{\boldsymbol{h}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{h}}_n\Big) + \boldsymbol{F}_{n+1}(\boldsymbol{s}^{p}_{n+1}) \qquad (10.56)$$

[4] It can be shown that the $h-$based formulation of the Picard method in form of (10.55) deduces from the more general $h-s-$based formulation of the Picard method in form of (10.39) if the saturation terms on the RHS of (10.39) are expressed by their derivatives with respect to the hydraulic head, viz.,

$$\boldsymbol{s}^{\tau+1}_{n+1} - \boldsymbol{s}_n - (1-\theta)\Delta t_n\dot{\boldsymbol{s}}_n = \boldsymbol{C}^{\tau}_{n+1}\cdot\big[\boldsymbol{h}^{\tau}_{n+1} - \boldsymbol{h}_n - (1-\theta)\Delta t_n\dot{\boldsymbol{h}}_n\big]$$

so that the storage matrix $\boldsymbol{O}^{\dagger}$ of the $h-$form results in

$$\boldsymbol{O}^{\dagger}(\boldsymbol{s}^{\tau}_{n+1}) = \boldsymbol{O}(\boldsymbol{s}^{\tau}_{n+1}) + \boldsymbol{B}\cdot\boldsymbol{C}^{\tau}_{n+1}$$

where the matrices $\boldsymbol{O}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are given from the $h-s-$form by (10.32) and (10.38), respectively.

by using the predictor solution $\boldsymbol{s}^p_{n+1}$ according to (10.43). A similar iterative procedure results for the θ−method (10.54). The iterations are terminated and the step-size error criteria are determined similar to (10.40) or (10.41) and (10.46), respectively.

In contrast to the Picard method, the full Newton method (cf. Sect. 8.18.2) can enhance the overall solution performance due to its quadratic convergence behavior. However, the resulting matrix system becomes unsymmetric. For the case of the GLS predictor-corrector time integration we find the following one-step Newton iteration scheme:

$$\Big(\frac{\boldsymbol{O}(\boldsymbol{s}^p_{n+1}) + \boldsymbol{B}\cdot\boldsymbol{C}^p_{n+1}}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}^p_{n+1}) + \hat{\boldsymbol{J}}(\boldsymbol{s}^p_{n+1})\Big)\cdot\boldsymbol{h}_{n+1} = \hat{\boldsymbol{J}}(\boldsymbol{s}^p_{n+1})\cdot\boldsymbol{h}^p_{n+1} + \\ \Big(\boldsymbol{O}(\boldsymbol{s}^p_{n+1}) + \boldsymbol{B}\cdot\boldsymbol{C}^p_{n+1}\Big)\cdot\Big(\frac{\boldsymbol{h}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{h}}_n\Big) + \boldsymbol{F}_{n+1}(\boldsymbol{s}^p_{n+1}) \qquad (10.57)$$

with the partial Jacobian

$$\hat{\boldsymbol{J}}(\boldsymbol{s}^p_{n+1}) = \frac{\partial\boldsymbol{O}(\boldsymbol{s}^p_{n+1})}{\partial\boldsymbol{h}^p_{n+1}}\cdot\Big(\frac{\boldsymbol{h}^p_{n+1}-\boldsymbol{h}_n}{\theta\Delta t_n} - \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{h}}_n\Big) + \\ \boldsymbol{h}^p_{n+1}\cdot\frac{\partial\boldsymbol{D}(\boldsymbol{s}^p_{n+1})}{\partial\boldsymbol{h}^p_{n+1}} - \frac{\partial\boldsymbol{F}(\boldsymbol{s}^p_{n+1})}{\partial\boldsymbol{h}^p_{n+1}} \qquad (10.58)$$

in which we have taken into account that $\boldsymbol{O}^\dagger(\boldsymbol{s}^p_{n+1}) = \boldsymbol{O}(\boldsymbol{s}^p_{n+1}) + \boldsymbol{B}\cdot\boldsymbol{C}^p_{n+1}$, where the matrices $\boldsymbol{O}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are defined in (10.32) and (10.38). The Newton scheme (10.57) can be recognized as a specific formulation within the more general PVST by choosing the hydraulic head h (equivalent to the pressure head ψ) as primary variable, which will be thoroughly described next in Sect. 10.7. The elements of the corresponding Jacobian (10.58) are derived in Appendix J for the equivalent ψ−based evaluation of the Jacobian $\boldsymbol{J}^\psi$.

10.7 Primary Variable Switching Technique (PVST)

Forsyth et al. [167] have introduced a powerful method in the context of unsaturated-saturated flow simulations, which is termed as the primary variable substitution, or *primary variable switching technique* (PVST). It originates from multiphase flow modeling and effectively handles the appearance and disappearance of phases [407]. 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 [141]. We will show the generality of

PVST for the class of unsaturated-saturated flow problems, which covers other solutions techniques as special cases, such as Celia et al.'s approximation method (Sect. 10.5.3). On the other hand, PVST is able to solve the Richards' equation in its $s-$form for heterogeneous media and overcomes the restrictions of common $s-$based solution methods as discussed in Sects. 10.3 and 10.4.

10.7.1 Basic Matrix System in the $\psi - s-$Formulation

The development is based on the mixed $\psi - s-$form of Richards' equation (10.5). Similar to the finite element formulation for the $h-s-$form of Richards' equation as given in Sect. 10.5 we find for the $\psi - s-$form of Richards' equation the following finite element matrix system

$$\boldsymbol{O}(\boldsymbol{s}) \cdot \dot{\boldsymbol{\psi}} + \boldsymbol{B} \cdot \dot{\boldsymbol{s}} + \boldsymbol{D}(\boldsymbol{s}) \cdot \boldsymbol{\psi} - \boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0} \tag{10.59}$$

where

$$\boldsymbol{\psi} = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{N_\mathrm{P}} \end{pmatrix}, \quad \boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{N_\mathrm{P}} \end{pmatrix}, \quad \dot{\boldsymbol{\psi}} = \begin{pmatrix} \frac{d\psi_1}{dt} \\ \frac{d\psi_2}{dt} \\ \vdots \\ \frac{d\psi_{N_\mathrm{P}}}{dt} \end{pmatrix}, \quad \dot{\boldsymbol{s}} = \begin{pmatrix} \frac{ds_1}{dt} \\ \frac{ds_2}{dt} \\ \vdots \\ \frac{ds_{N_\mathrm{P}}}{dt} \end{pmatrix} \tag{10.60}$$

and

$$\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \delta_{ij} \int_{\Omega^e} s^e(\psi^e) S_o^e \, N_i \, d\Omega^e \\
\boldsymbol{B} = B_{ij} &= \sum_e \delta_{ij} \int_{\Omega^e} \varepsilon^e \, N_i \, d\Omega^e \\
\boldsymbol{D} = D_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot \big(k_r^e(s^e) \boldsymbol{K}^e f_\mu^e \cdot \nabla N_j \big) d\Omega^e + \int_{\Gamma_C^e} \Phi_h^e N_i N_j \, d\Gamma^e \Big) \\
\boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i (Q_h^e + Q_{\mathrm{EOB}}^e) d\Omega^e - \\
&\quad \int_{\Omega^e} \nabla N_i \cdot \big(k_r^e(s^e) \boldsymbol{K}^e f_\mu^e \cdot (1 + \chi^e) \boldsymbol{e} \big) d\Omega^e + \\
&\quad \int_{\Gamma_C^e} N_i \Phi_h^e h_C^e \, d\Gamma^e - \int_{\Gamma_N^e} N_i q_h^e \, d\Gamma^e - \int_{\Gamma_N^{\nabla e}} N_i k_r^e(s^e) q_h^{\nabla e} \, d\Gamma^e \Big) - \\
&\quad Q_w(t) \Big|_i
\end{aligned} \tag{10.61}$$

which has to be solved for the discrete pressure head variable $\boldsymbol{\psi}$ and discrete saturation variable $\boldsymbol{s}$. Using implicit methods of time integration, (10.59) reads

$$\boldsymbol{O}(\boldsymbol{s}_{n+1}) \cdot \dot{\boldsymbol{\psi}}_{n+1} + \boldsymbol{B} \cdot \dot{\boldsymbol{s}}_{n+1} + \boldsymbol{D}(\boldsymbol{s}_{n+1}) \cdot \boldsymbol{\psi}_{n+1} - \boldsymbol{F}(\boldsymbol{s}_{n+1}) = \boldsymbol{0} \tag{10.62}$$

where the time derivatives are approximated by

$$\dot{\boldsymbol{\psi}}_{n+1} = \frac{\boldsymbol{\psi}_{n+1} - \boldsymbol{\psi}_n}{\theta \Delta t_n} - \left(\tfrac{1}{\theta} - 1\right)\dot{\boldsymbol{\psi}}_n, \quad \dot{\boldsymbol{s}}_{n+1} = \frac{\boldsymbol{s}_{n+1} - \boldsymbol{s}_n}{\theta \Delta t_n} - \left(\tfrac{1}{\theta} - 1\right)\dot{\boldsymbol{s}}_n \tag{10.63}$$

in which the weighting factor $\theta \in (\frac{1}{2}, 1)$ is $\frac{1}{2}$ the TR scheme and unity for the BE scheme. Inserting (10.63) into (10.62) results in

$$\begin{aligned}\boldsymbol{R}_{n+1}(\boldsymbol{\psi}, \boldsymbol{s}) = &\left(\frac{\boldsymbol{O}(\boldsymbol{s}_{n+1})}{\theta \Delta t_n} + \boldsymbol{D}(\boldsymbol{s}_{n+1})\right) \cdot \boldsymbol{\psi}_{n+1} + \frac{\boldsymbol{B}}{\theta \Delta t_n} \cdot \boldsymbol{s}_{n+1} - \\ &\boldsymbol{O}(\boldsymbol{s}_{n+1}) \cdot \left(\frac{\boldsymbol{\psi}_n}{\theta \Delta t_n} + \left(\tfrac{1}{\theta} - 1\right)\dot{\boldsymbol{\psi}}_n\right) - \\ &\boldsymbol{B} \cdot \left(\frac{\boldsymbol{s}_n}{\theta \Delta t_n} + \left(\tfrac{1}{\theta} - 1\right)\dot{\boldsymbol{s}}_n\right) - \boldsymbol{F}(\boldsymbol{s}_{n+1}) = \boldsymbol{0}\end{aligned} \tag{10.64}$$

where $\boldsymbol{R}$ represents the *residual* of the spatio-temporarily discretized $\psi - s-$form of Richards' equation.

10.7.2 Primary Variable Switching Methodology

To solve the basic matrix system (10.64) one has to decide which variable of ψ or s should be primary. Commonly, the selection of the primary variable is done in a static manner and results in a 'fixed' $\psi-$, $s-$ or $\psi - s-$modeling strategy, including the limitations and drawbacks discussed above in Sect. 10.3. In contrast, primary variable switching is done dynamically depending on the current flow characteristics.

Let X_i be the primary variable associated with the global node i. X_i can be either ψ_i or s_i. Accordingly, we can consider $\boldsymbol{X}$ as a nodal vector containing the different primary variables in the solution space Ω as

$$\boldsymbol{X} \in (\boldsymbol{\psi}, \boldsymbol{s}) \tag{10.65}$$

Hence, the matrix system (10.64) can be written in the form

$$\boldsymbol{R}_{n+1}(\boldsymbol{X}) = \boldsymbol{0} \tag{10.66}$$

and solved for X_i $(i = 1, \ldots, N_{\mathrm{P}})$. The solution of the nonlinear equations (10.66), i.e., the vector of primary variables $\boldsymbol{X}$, is performed by the Newton method (cf. Sect. 8.18.2), viz.,

$$\boldsymbol{J}^X(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})\Delta \boldsymbol{X}_{n+1}^{\tau} = -\boldsymbol{R}_{n+1}^{\tau}(\boldsymbol{\psi}, \boldsymbol{s}) \tag{10.67}$$

with the increment

$$\Delta \boldsymbol{X}_{n+1}^{\tau} = \boldsymbol{X}_{n+1}^{\tau+1} - \boldsymbol{X}_{n+1}^{\tau} \tag{10.68}$$

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

$$J_{ij}^X(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau}) = \frac{\partial R_{i,n+1}(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})}{\partial X_{j,n+1}^{\tau}} \tag{10.69}$$

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 [167]:

```
IF ($s_{i,n+1}^{\tau} \geq \mathrm{tol}_f$) THEN
    Use $\psi_{i,n+1}^{\tau}$ as primary variable at node $i$ and solve the Newton statement (10.67)
    as
```

$$J_{ij}^{\psi}(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})\Delta \psi_{j,n+1}^{\tau} = -R_{i,n+1}^{\tau}(\boldsymbol{\psi}, \boldsymbol{s}) \tag{10.70}$$

```
ELSE IF ($s_{i,n+1}^{\tau} < \mathrm{tol}_b$) THEN
    Use $s_{i,n+1}^{\tau}$ as primary variable at node $i$ and solve the Newton statement
    (10.67) as
```

$$J_{ij}^{s}(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})\Delta s_{j,n+1}^{\tau} = -R_{i,n+1}^{\tau}(\boldsymbol{\psi}, \boldsymbol{s}) \tag{10.71}$$

```
ELSE
    Do not change primary variable for the node $i$ and solve (10.70) or (10.71)
    according to the hitherto selected primary variable ($\psi_{i,n+1}^{\tau}$ or $s_{i,n+1}^{\tau}$).
ENDIF
```

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

$$\mathrm{tol}_f < 1, \quad \mathrm{tol}_f \neq \mathrm{tol}_b \tag{10.72}$$

Useful switching tolerances are [141, 167]

$$\mathrm{tol}_f = 0.99, \quad \mathrm{tol}_b = 0.89 \tag{10.73}$$

The Newton approach requires continuous derivatives of the Jacobians $\boldsymbol{J}^{\psi}$ and $\boldsymbol{J}^{s}$ with respect to the pressure head ψ and the saturation s, respectively. In the present FEM the variables ψ and s are approximated in a continuous manner according to (10.27) if occurring as primary variables and the Jacobians are thus derivable. The Jacobians $\boldsymbol{J}^X$ can be computed either numerically or analytically (cf. Sect. 8.18.2). The analytical method has shown more efficient [335] and will be preferred in the following. While a perturbation scheme such as the one used by Forsyth et al. [167] requires a pass of $2N_\text{P}$ evaluations, analytic derivatives require only a pass of N_P evaluations. The elements of the corresponding Jacobians $J_{ij}^{\psi}(\psi_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})$ of (10.70) and $J_{ij}^{s}(\psi_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})$ of (10.71) are summarized in the Appendix J. Otherwise, the residual $R_{i,n+1}^{\tau}(\psi, \boldsymbol{s})$ at the iterate τ and node i is independent of the actually used primary variables X_i and is computed according to (10.64) in the following way

$$\begin{aligned} -R_{i,n+1}^{\tau}(\psi, \boldsymbol{s}) = &-\left(\frac{O_{ij}(\boldsymbol{s}_{n+1}^{\tau})}{\theta \Delta t_n} + D_{ij}(\boldsymbol{s}_{n+1}^{\tau})\right)\psi_{j,n+1}^{\tau} - \frac{B_{ij}}{\theta \Delta t_n}\, s_{j,n+1}^{\tau} + \\ &O_{ij}(\boldsymbol{s}_{n+1}^{\tau})\left(\frac{\psi_{j,n}}{\theta \Delta t_n} + \left(\tfrac{1}{\theta} - 1\right)\dot{\psi}_{j,n}\right) + \\ &B_{ij}\left(\frac{s_{j,n}}{\theta \Delta t_n} + \left(\tfrac{1}{\theta} - 1\right)\dot{s}_{j,n}\right) + F_i(\boldsymbol{s}_{n+1}^{\tau}) \end{aligned} \qquad (10.74)$$

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

$$J_{ij}^X = \sum_e \int_{\Omega^e} (\ldots)_{\forall i, \forall j} \qquad (10.75)$$

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

$$J_{ij}^X = \sum_i \sum_{e \in \eta_i} \int_{\Omega^e} (\ldots)_{i, \forall j} \qquad (10.76)$$

where the contributions from an adjacent element patch η_i to a node i are added in the global matrix.

10.7.3 Deducing Standard Schemes from PVST

PVST can be considered as a general formulation in which standard solution strategies are comprised as special cases. We choose the pressure head ψ as primary variable and start from the Newton statement (10.70) written in the form

$$\boldsymbol{J}^{\psi}(\psi_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau}) \cdot (\psi_{n+1}^{\tau+1} - \psi_{n+1}^{\tau}) = -\boldsymbol{R}_{n+1}^{\tau}(\psi, \boldsymbol{s}) \qquad (10.77)$$

for solving the pressure head $\psi_{n+1}^{\tau+1}$ at new iterate $\tau+1$ and new time plane $n+1$, where the Jacobian $\boldsymbol{J}^{\psi}$ is given by (J.1) of Appendix J and the residual is defined by (10.64).

10.7.3.1 Newton Scheme of Mixed $\psi - s$–Formulation

Inserting the partial Jacobians of $\boldsymbol{J}^{\psi}$ derived in Appendix J and the residual $\boldsymbol{R}$ of (10.64) in (10.77) the following Newton scheme of the mixed $\psi - s$–form of the Richards' equation results:

$$\begin{aligned}\Big(\frac{\boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau}) + \boldsymbol{B}\cdot\boldsymbol{C}_{n+1}^{\tau}}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}_{n+1}^{\tau}) + \hat{\boldsymbol{J}}(\boldsymbol{s}_{n+1}^{\tau})\Big)\cdot\boldsymbol{\psi}_{n+1}^{\tau+1} = \\ \boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau})\cdot\Big(\frac{\boldsymbol{\psi}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{\psi}}_n\Big) + \Big(\frac{\boldsymbol{B}}{\theta\Delta t_n}\cdot\boldsymbol{C}_{n+1}^{\tau} + \hat{\boldsymbol{J}}(\boldsymbol{s}_{n+1}^{\tau})\Big)\cdot\boldsymbol{\psi}_{n+1}^{\tau} + \\ \boldsymbol{B}\cdot\Big(\big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{s}}_n - \frac{\boldsymbol{s}_{n+1}^{\tau}-\boldsymbol{s}_n}{\theta\Delta t_n}\Big) + \boldsymbol{F}_{n+1}(\boldsymbol{s}_{n+1}^{\tau})\end{aligned} \tag{10.78}$$

with the partial Jacobian

$$\begin{aligned}\hat{\boldsymbol{J}}(\boldsymbol{s}_{n+1}^{\tau}) &= \frac{\partial\boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau})}{\partial\boldsymbol{\psi}_{n+1}^{\tau}}\cdot\Big(\frac{\boldsymbol{\psi}_{n+1}^{\tau}-\boldsymbol{\psi}_n}{\theta\Delta t_n} - \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{\psi}}_n\Big) + \\ &\qquad \boldsymbol{\psi}_{n+1}^{\tau}\cdot\frac{\partial\boldsymbol{D}(\boldsymbol{s}_{n+1}^{\tau})}{\partial\boldsymbol{\psi}_{n+1}^{\tau}} - \frac{\partial\boldsymbol{F}(\boldsymbol{s}_{n+1}^{\tau})}{\partial\boldsymbol{\psi}_{n+1}^{\tau}} \\ &= \boldsymbol{J}^{\psi 3} + \boldsymbol{J}^{\psi 4} - \boldsymbol{J}^{\psi 5}\end{aligned} \tag{10.79}$$

which elements are evaluated in Sect. J.1 of Appendix J in form of $\boldsymbol{J}^{\psi 3}$ (J.4), $\boldsymbol{J}^{\psi 4}$ (J.5) and $\boldsymbol{J}^{\psi 5}$ (J.6). Due to the partial Jacobian $\hat{\boldsymbol{J}}$ the resulting matrix system (10.78) is unsymmetric.

10.7.3.2 Picard Scheme of Mixed $\psi - s$–Formulation: Celia et al.'s Approximation Method

If we drop the partial Jacobian $\hat{\boldsymbol{J}}$ in (10.78) the matrix system reduces to

$$\begin{aligned}\Big(\frac{\boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau}) + \boldsymbol{B}\cdot\boldsymbol{C}_{n+1}^{\tau}}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}_{n+1}^{\tau})\Big)\cdot\boldsymbol{\psi}_{n+1}^{\tau+1} = \\ \boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau})\cdot\Big(\frac{\boldsymbol{\psi}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{\psi}}_n\Big) + \Big(\frac{\boldsymbol{B}}{\theta\Delta t_n}\cdot\boldsymbol{C}_{n+1}^{\tau}\Big)\cdot\boldsymbol{\psi}_{n+1}^{\tau} + \\ \boldsymbol{B}\cdot\Big(\big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{s}}_n - \frac{\boldsymbol{s}_{n+1}^{\tau}-\boldsymbol{s}_n}{\theta\Delta t_n}\Big) + \boldsymbol{F}_{n+1}(\boldsymbol{s}_{n+1}^{\tau})\end{aligned} \tag{10.80}$$

which can be recognized as Celia et al.'s approximation method of Picard iteration equivalent to (10.39) derived in Sect. 10.5.3 for the mixed $h-s-$form of Richards' equation. The resulting matrix system (10.80) is symmetric.

10.7.3.3 Newton Scheme of Standard $\psi-$Formulation

If the saturation terms are expressed by derivatives with respect to the pressure head in the form

$$\boldsymbol{s}_{n+1}^{\tau+1} - \boldsymbol{s}_n - (1-\theta)\Delta t_n \dot{\boldsymbol{s}}_n = \boldsymbol{C}_{n+1}^{\tau} \cdot \left[\boldsymbol{\psi}_{n+1}^{\tau} - \boldsymbol{\psi}_n - (1-\theta)\Delta t_n \dot{\boldsymbol{\psi}}_n\right] \tag{10.81}$$

we find from (10.78) the Newton scheme of the standard $\psi-$formulation, viz.,

$$\begin{aligned} \Big(\frac{\boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau}) + \boldsymbol{B}\cdot\boldsymbol{C}_{n+1}^{\tau}}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}_{n+1}^{\tau}) + \hat{\boldsymbol{J}}(\boldsymbol{s}_{n+1}^{\tau})\Big)\cdot\boldsymbol{\psi}_{n+1}^{\tau+1} = \\ \Big(\boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau}) + \boldsymbol{B}\cdot\boldsymbol{C}_{n+1}^{\tau}\Big)\cdot\Big(\frac{\boldsymbol{\psi}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{\psi}}_n\Big) + \hat{\boldsymbol{J}}(\boldsymbol{s}_{n+1}^{\tau})\cdot\boldsymbol{\psi}_{n+1}^{\tau} + \\ \boldsymbol{F}_{n+1}(\boldsymbol{s}_{n+1}^{\tau}) \end{aligned} \tag{10.82}$$

where the partial Jacobian $\hat{\boldsymbol{J}}$ is defined by (10.79). The matrix system is unsymmetric. We recognize that (10.82) of the $\psi-$formulation is equivalent to (10.57) of a $h-$formulation derived in Sect. 10.6.2.

10.7.3.4 Picard Scheme of Standard $\psi-$Formulation

Dropping the partial Jacobian $\hat{\boldsymbol{J}}$ in (10.82) the Picard scheme of the standard $\psi-$formulation results:

$$\begin{aligned} \Big(\frac{\boldsymbol{O}^{\dagger}(\boldsymbol{s}_{n+1}^{\tau})}{\theta\Delta t_n} + \boldsymbol{D}(\boldsymbol{s}_{n+1}^{\tau})\Big)\cdot\boldsymbol{\psi}_{n+1}^{\tau+1} = \boldsymbol{O}^{\dagger}(\boldsymbol{s}_{n+1}^{\tau})\cdot\Big(\frac{\boldsymbol{\psi}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{\psi}}_n\Big) + \\ \boldsymbol{F}_{n+1}(\boldsymbol{s}_{n+1}^{\tau}) \end{aligned} \tag{10.83}$$

where

$$\boldsymbol{O}^{\dagger}(\boldsymbol{s}_{n+1}^{\tau}) = \boldsymbol{O}(\boldsymbol{s}_{n+1}^{\tau}) + \boldsymbol{B}\cdot\boldsymbol{C}_{n+1}^{\tau} \tag{10.84}$$

which is shown equivalent to (10.55) derived in Sect. 10.6.2 for the $h-$form of Richards' equation. The resulting matrix system is symmetric.

10.7.4 Convergence Criteria

An important aspect of the iterative solution of the nonlinear system (10.67) is the choice of appropriate convergence criteria. Commonly, as a standard test, the deviatory (change) convergence criterion in a form of (8.378) is applied to terminate the Newton (or Picard) iterations [156]. For the GLS predictor-corrector time integration method a deviatory error measure $\|\boldsymbol{d}_{n+1}\|$ as a function of the difference between corrector and predictor solution $(\boldsymbol{X}_{n+1} - \boldsymbol{X}_{n+1}^{p})$ controls the time step lengths via a user-specified dimensionless tolerance ϵ (cf. Table 8.7). In a one-step Newton (or one-step Picard) iteration method the deviatory error measure even implicitly controls the iteration error $(\boldsymbol{X}_{n+1}^{\tau+1} - \boldsymbol{X}_{n+1}^{\tau}),\ \tau = 0, 1$ by the same tolerance ϵ too.

For the present problem class of unsaturated-saturated flow the control of only the deviatory error can be insufficient because a small change in the primary variable can still implicate notable mass defects in the governing discrete balance equations. For this reason an additional direct control of the residual $\boldsymbol{R}_{n+1}^{\tau} \to \boldsymbol{0}$ in (10.67) can be appropriate. It provides a direct measurement of the global mass balance error after terminating the Newton iteration. For instance one can enforce the condition

$$\|\boldsymbol{R}_{n+1}^{\tau}\| \leq \epsilon_2 \|\boldsymbol{F}_{n+1}\| \tag{10.85}$$

where a second dimensionless tolerance ϵ_2 is introduced and an appropriate normalization of the residual (here with respect to the external supply $\boldsymbol{F}_{n+1}$) is applied. However, since $\boldsymbol{F}_{n+1}$ is also nonlinearly dependent on the solution it has been shown often more robust to control the residual without normalization, viz.,

$$\|\boldsymbol{R}_{n+1}^{\tau}\| \leq \epsilon_2^{*} \tag{10.86}$$

where ϵ_2^{*} is a dimensional residual error tolerance. In (10.85) and (10.86) $\|.\|$ corresponds to a suitable error norm, e.g., RMS error norm (8.28) or maximum error norm (8.29). While the residual control can also be used as an exclusive convergence criterion [167], we usually prefer the control of both the deviatory and the residual errors. It has shown the best iteration strategy to minimize both temporal truncation and mass balance errors [141]. In doing so, the matrix system has to be solved in the basic matrix form of (10.67) to directly evaluate the residual $\boldsymbol{R}_{n+1}^{\tau}$ during the iterations τ. This requires little extra work compared to a common non-residual-written matrix form such as used in (10.78). It is important to note that any residual control does not allow anymore a one-step Newton (or one-step Picard) iteration within the GLS predictor-corrector time marching strategy. The different options available in the overall solution control of the adaptive GLS predictor-corrector method will be described in Sect. 10.7.5.

To measure the 'accumulated loss' of mass over an entire simulation period (t_0, t_{end}) we can record the *total balance error* $\text{TBE}(t_{\text{end}})$ defined as

$$\mathrm{TBE}(t_{\mathrm{end}}) = \int_{t=t_0}^{t_{\mathrm{end}}} \|\boldsymbol{R}^{\tau}(t)\| dt \tag{10.87}$$

where $\|.\|$ is a proper (RMS or maximum) error norm.

10.7.5 Solution Control

Generally, the control of the solution of the resulting highly nonlinear matrix systems is a tricky matter. Both the choice of the time step size Δt_n and the iteration control of the iteration 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 in form of the GLS predictor-corrector time integrator has shown most useful for the present class of problems [141]. 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 PVST the Newton method plays a central role. 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.

For PVST the overall iterative solution method embedded in the GLS predictor-corrector time marching strategy consists of the following main working steps:

STEP 0: Initialization

Compute the initial acceleration vectors $\dot{\boldsymbol{\psi}}_0$ and $\dot{\boldsymbol{s}}_0$ from (10.62) as

$$\big(\boldsymbol{O}(\boldsymbol{s}_0) + \boldsymbol{B} \cdot \boldsymbol{C}_0\big) \cdot \dot{\boldsymbol{\psi}}_0 = -\boldsymbol{D}(\boldsymbol{s}_0) \cdot \boldsymbol{\psi}_0 + \boldsymbol{F}(\boldsymbol{s}_0) \tag{10.88}$$

and with

$$\dot{\boldsymbol{s}}_0 = \boldsymbol{C}_0 \cdot \dot{\boldsymbol{\psi}}_0 \tag{10.89}$$

where $\boldsymbol{C}_0$ is the moisture capacity matrix (10.38) evaluated at initial time t_0, $\boldsymbol{\psi}_0$ and $\boldsymbol{s}_0$ are the initial distributions of the pressure head ψ and the saturation s, respectively. Furthermore, we choose a small initial time step size Δt_0.

STEP 1: Predictor solutions

Explicit schemes of 1st-order and 2nd-order accuracy in time provide appropriate predictor solutions for the primary variable $\boldsymbol{X}_{n+1}$ (either ψ_{n+1} or $\boldsymbol{s}_{n+1}$) at the new time plane $n+1$. We use either the 1st-order accurate FE and 2nd-order accurate AB scheme, respectively,

$$\boldsymbol{X}_{n+1}^{p} = \begin{cases} \boldsymbol{X}_n + \Delta t_n \dot{\boldsymbol{X}}_n & \text{FE predictor} \\ \boldsymbol{X}_n + \frac{\Delta t_n}{2}\big[\big(2 + \frac{\Delta t_n}{\Delta t_{n-1}}\big)\dot{\boldsymbol{X}}_n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\boldsymbol{X}}_{n-1}\big] & \text{AB predictor} \end{cases} \tag{10.90}$$

Note here that, since $\dot{\boldsymbol{X}}_{n-1}$ is required, the AB formula cannot be applied before the second step ($n=1$). The prediction has to be started with the FE procedure, where $\dot{\boldsymbol{X}}_0$ is available from (10.88) and (10.89). The superscript p indicates the predictor values at the new time plane $n+1$. To initialize the iteration procedure for $\tau = 0$ we take

$$\boldsymbol{\psi}_{n+1}^{0} = \boldsymbol{\psi}_{n+1}^{p}, \quad \boldsymbol{s}_{n+1}^{0} = \boldsymbol{s}_{n+1}^{p} \tag{10.91}$$

STEP 2: Corrector solutions

(i) *PVST with residual control:* Depending on the primary variable switching criteria stated above the following iteration procedure of the matrix systems (10.70), (10.71) arises to solve the pressure head $\boldsymbol{\psi}_{n+1}$ or the saturation $\boldsymbol{s}_{n+1}$, respectively,

```
For iterations $\tau = 0, 1, 2, \ldots$ compute until convergence:
    For each node $i = 1, \ldots, N_{\mathrm{P}}$ do either:
        $J_{ij}^{\psi}(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})\Delta\psi_{j,n+1}^{\tau} = -R_{i,n+1}^{\tau}(\boldsymbol{\psi}, \boldsymbol{s})$
        $\Delta\psi_{j,n+1}^{\tau} = \psi_{j,n+1}^{\tau+1} - \psi_{j,n+1}^{\tau}$
    or                                                                    (10.92)
        $J_{ij}^{s}(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})\Delta s_{j,n+1}^{\tau} = -R_{i,n+1}^{\tau}(\boldsymbol{\psi}, \boldsymbol{s})$
        $\Delta s_{j,n+1}^{\tau} = s_{j,n+1}^{\tau+1} - s_{j,n+1}^{\tau}$
    End do
Stop if $\|\boldsymbol{R}_{n+1}^{\tau}(\boldsymbol{\psi}, \boldsymbol{s})\|_{\mathrm{RMS}} \le \epsilon_2^{*}$
To obtain the corrector solutions: $\boldsymbol{\psi}_{n+1} = \boldsymbol{\psi}_{n+1}^{\tau+1}$, $\boldsymbol{s}_{n+1} = \boldsymbol{s}_{n+1}^{\tau+1}$
```

Note that the predictor of the FE (10.90) is used for the BE ($\theta = 1$) and that the predictor of the AB (10.90) is used for the TR ($\theta = \frac{1}{2}$) in (10.92). Accordingly, the predictor-corrector solutions are called FE/BE and AB/TR scheme, respectively.

(ii) *PVST of one-step Newton method without residual control:* Note that the additional residual test (10.86) can be optionally omitted in (10.92). In this case the one-step Newton method (i.e., $\tau = 1$) is used and the deviatory error criterion ϵ of the GLS predictor-corrector method also controls the convergence of the Newton iteration via the adaptive time steps. The corrector solutions are immediately obtained via

$$\begin{aligned}
&\textit{For each node } i = 1, \ldots, N_{\mathrm{P}} \textit{ do either:}\\
&\qquad J^{\psi}_{ij}(\boldsymbol{\psi}^{p}_{n+1}, \boldsymbol{s}^{p}_{n+1})\Delta\psi_{j,n+1} = -R^{p}_{i,n+1}(\boldsymbol{\psi}, \boldsymbol{s})\\
&\qquad \Delta\psi_{j,n+1} = \psi_{j,n+1} - \psi^{p}_{j,n+1}\\
&\textit{or}\\
&\qquad J^{s}_{ij}(\boldsymbol{\psi}^{p}_{n+1}, \boldsymbol{s}^{p}_{n+1})\Delta s_{j,n+1} = -R^{p}_{i,n+1}(\boldsymbol{\psi}, \boldsymbol{s})\\
&\qquad \Delta s_{j,n+1} = s_{j,n+1} - s^{p}_{j,n+1}\\
&\textit{End do}
\end{aligned} \tag{10.93}$$

where the predictor solutions $\boldsymbol{\psi}^{p}_{n+1}$ and $\boldsymbol{s}^{p}_{n+1}$ are used to linearize $\boldsymbol{J}^{\psi}$, $\boldsymbol{J}^{s}$ and $\boldsymbol{R}_{n+1}$.

(iii) *Newton or Picard iteration with residual control at enforced ψ−variable (suppressed PVST):* Optionally, it can be useful to suppress variable switching and solve the matrix system always in the pressure head variable $\boldsymbol{\psi}$, viz.,

$$\begin{aligned}
&\textit{For iterations } \tau = 0, 1, 2, \ldots \textit{ compute until convergence:}\\
&\qquad \boldsymbol{J}^{\psi}(\boldsymbol{\psi}^{\tau}_{n+1}, \boldsymbol{s}^{\tau}_{n+1}) \cdot \Delta\boldsymbol{\psi}^{\tau}_{n+1} = -\boldsymbol{R}^{\tau}_{n+1}(\boldsymbol{\psi}, \boldsymbol{s})\\
&\qquad \Delta\boldsymbol{\psi}^{\tau}_{n+1} = \boldsymbol{\psi}^{\tau+1}_{n+1} - \boldsymbol{\psi}^{\tau}_{n+1}\\
&\textit{Stop if } \ \|\boldsymbol{R}^{\tau}_{n+1}(\boldsymbol{\psi}, \boldsymbol{s})\|_{\mathrm{RMS}} \le \epsilon^{*}_{2}\\
&\textit{To obtain the corrector solutions: } \boldsymbol{\psi}_{n+1} = \boldsymbol{\psi}^{\tau+1}_{n+1},\ \boldsymbol{s}_{n+1} = f(\boldsymbol{\psi}^{\tau+1}_{n+1})
\end{aligned} \tag{10.94}$$

The matrix solution (10.94) can be applied to both Newton and Picard iteration. The Picard method can be sometimes favorable due to its higher robustness and preserving matrix symmetry, albeit its lower convergence rate, cf. Sect. 8.18.1. For running the Picard iteration in (10.94) the derivative terms in the Jacobian $\boldsymbol{J}^{\psi}$ are dropped, so that $\boldsymbol{J}^{\psi} = \boldsymbol{J}^{\psi 1} + \boldsymbol{J}^{\psi 2}$ according to (J.1) of Appendix J.

STEP 3: Updated accelerations

In preparing the data for the next time step the new acceleration vectors $\dot{\boldsymbol{X}}_{n+1}$ are computed according to Table 8.7 as

$$\dot{\boldsymbol{X}}_{n+1} = \begin{cases} \frac{\boldsymbol{X}_{n+1}-\boldsymbol{X}_{n}}{\Delta t_n} & \text{FE}\\ \big(2 - \frac{\Delta t_{n-1}}{\Delta t_n + \Delta t_{n-1}}\big)\big(\frac{\boldsymbol{X}_{n+1}-\boldsymbol{X}_{n}}{\Delta t_n}\big) - \big(\frac{\Delta t_n}{\Delta t_n + \Delta t_{n-1}}\big)\big(\frac{\boldsymbol{X}_{n}-\boldsymbol{X}_{n-1}}{\Delta t_{n-1}}\big) & \text{AB} \end{cases} \tag{10.95}$$

STEP 4: Error estimation

The LTE of the approximate equations depends on the predicted $\boldsymbol{X}^{p}_{n+1}$ and corrected $\boldsymbol{X}_{n+1}$ solutions. For the FE/BE and the AB/TR the error estimation yields (cf. Table 8.7)

$$\boldsymbol{d}_{n+1} = \varphi(\boldsymbol{X}_{n+1} - \boldsymbol{X}^{p}_{n+1}) \tag{10.96}$$

with

$$\varphi = \begin{cases} \frac{1}{2} & \text{for FE/BE} \\ \frac{1}{3\left(1+\frac{\Delta t_{n-1}}{\Delta t_n}\right)} & \text{for AB/TR} \end{cases} \tag{10.97}$$

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

$$\|\boldsymbol{d}_{n+1}\|_{L_2} = \left[\frac{1}{N_\mathrm{P}}\left(\sum_{i=1}^{N_\mathrm{P}}\left|\frac{d_{i,n+1}}{X_{\max,n+1}}\right|^2\right)\right]^{1/2} \tag{10.98}$$

and the maximum L_∞ error norm

$$\|\boldsymbol{d}_{n+1}\|_{L_\infty} = \frac{1}{X_{\max,n+1}} \max_i |d_{i,n+1}| \tag{10.99}$$

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 (10.96), (10.98), (10.99), the current time step size Δt_n, and a user-specified error tolerance ϵ as

$$\Delta t_{n+1} = \begin{cases} \Delta t_n\left(\frac{\epsilon}{\|\boldsymbol{d}_{n+1}\|_{L_p}}\right)^{1/\lambda} & \text{or optionally} \\ \Delta t_n\left(\frac{\epsilon}{\max(\|\boldsymbol{d}^h_{n+1}\|_{L_p}, \|\boldsymbol{d}^\psi_{n+1}\|_{L_p}, \|\boldsymbol{d}^s_{n+1}\|_{L_p})}\right)^{1/\lambda} & \end{cases} \tag{10.100}$$

where

$$\begin{aligned} \lambda &= \begin{cases} 2 & \text{for FE/BE} \\ 3 & \text{for AB/TR} \end{cases} \\ p &= \begin{cases} 2 & \text{for RMS error norm} \\ \infty & \text{for maximum error norm} \end{cases} \end{aligned} \tag{10.101}$$

and $\boldsymbol{d}^h_{n+1}$, $\boldsymbol{d}^\psi_{n+1}$ and $\boldsymbol{d}^s_{n+1}$ are defined by (10.45). To monitor the progress of the solution we use the criteria as summarized in Table 8.7.

Note that alternative time stepping schemes exist in contrast to the GLS predictor-corrector method, e.g., the empirical target-based scheme as proposed in [167], in which the only criterion is the Newton convergence for 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.[5] Diersch and Perrochet [141] have studied the target-based solution control and concluded that the GLS predictor-corrector method is usually superior to a target-based scheme, which can be an error-prone strategy in a potential lacking of temporal accuracy.

10.8 Overview of FEFLOW's Solution Strategies for Unsaturated-Saturated Flow

In the preceding sections a family of schemes has been developed for solving the Richards' equation in different formulations. The complexity and numerical difficulties which can arise in the practical solution of unsaturated-saturated flow actually require the availability of a spectrum of methods having their advantages and drawbacks. In Table 10.3 we summarize the solution strategies with their essential features and options available in FEFLOW.[6]

[5] *Empirical target-based time step control*: If Newton iterations have converged a new provisional step size Δt_{n+1} can be computed in the following way [141]:

$$\Delta t_{n+1} = \Xi \; \Delta t_n$$

where Ξ is a time step multiplier, which is determined by the minimum ratio of prescribed target change parameters DXWISH (DSWISH for the saturations $\boldsymbol{s}_{n+1}$ and DPWISH for the pressure head $\boldsymbol{\psi}_{n+1}$) to the Newton correction, viz.,

$$\Xi = \min_i \frac{\text{DXWISH}}{|X_{i,n+1}^{\tau+1} - X_{i,n}|}$$

Typically used values are DSWISH $= 0.4$ and DPWISH $= 400$ m. Additionally, it can be useful to constrain Ξ by a maximum multiplier $\Xi \leq \Xi_{\max}$, where $\Xi_{\max} = 1.1, \ldots, 5$. If the Newton scheme does not converge within a maximum number of iterations $\tau \leq$ ITMAX, where ITMAX is typically 12, the current time step has to be rejected. A reduced time step size is then computed by $\Delta t_n^{\text{red}} = \Delta t_n/\text{TDIV}$ and the solution process is restarted for the current time plane $n+1$, but with $\Delta t_n = \Delta t_n^{\text{red}}$. The time step divider TDIV is usually 2.

[6] Of primary interests are the schemes no.1, no.3 and no.4, providing a full residual control and best mass-conservative properties. Scheme no.1 is very effective for dry porous media, however, it is not well applicable to hysteretic porous-media problems. The Picard method of scheme no.4 is potentially more robust compared to the Newton scheme no.3, however, to the disadvantage of only a linear convergence rate. In solving the mixed $\psi - s-$form (or the equivalent $h - s-$form) of Richards' equation, the moisture capacity C is usually evaluated analytically. On the other hand, for the standard $h-$based forms of the Richards' equation the chord slope evaluation (see Sect. J.3 of Appendix J) of the moisture capacity is often preferred due to a potentially better discrete mass conservation property. The $h-$form of schemes no.9 and no.7 are suited for classic seepage simulations (at moderate capillary pressure conditions) involving free surface(s). Scheme no.9 with $\theta = 1$ can be used to approach to steady-state solutions (whenever exist).

Table 10.3 Schemes and options for solving the Richards' equation of unsaturated-saturated flow available in FEFLOW

						Options		
No.	Form	Method	Time stepping	Primary variable	Matrix system	Iteration method	Residual control	Moisture capacity
1	Mixed $\psi - s-$form	PVST	Predictor-corrector	ψ or s	(10.92)	Full Newton	Yes	Analytic[a]
2	Mixed $\psi - s-$form	PVST	Predictor-corrector	ψ or s	(10.93)	One-step Newton	No	Analytic[a]
3	Mixed $\psi - s-$form		Predictor-corrector	ψ	(10.94)	Full Newton	Yes	Analytic[a]
4	Mixed $\psi - s-$form		Predictor-corrector	ψ	(10.94)	Picard	Yes	Analytic[a]
5	Mixed $\psi - s-$form		Predictor-corrector	ψ	(10.78)	One-step Newton	No	Analytic[a]
6	Mixed $h - s-$form	Celia et al.	Predictor-corrector	h	(10.42)	One-step Picard	No	Analytic[a]
7	$h-$form	Standard	Predictor-corrector	h	(10.56)	One-step Picard	No	Chord slope[b]
8	$h-$form	Standard	Predictor-corrector	h	(10.57)	One-step Newton	No	Chord slope[b]
9	$h-$form	Standard	$\theta-$method	h	(10.55)	Picard	No	Chord slope[b]

[a] Optionally, chord slope and time-centered evaluations, cf. Sect. J.3 of Appendix J and Appendix D, respectively

[b] Optionally, analytic evaluations, cf. Appendix D

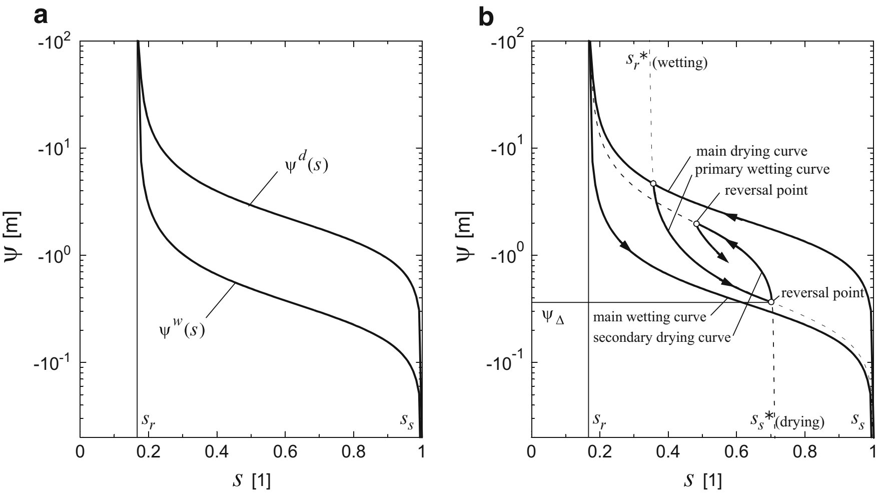


Fig. 10.4 Typical plots of (**a**) main $\psi(s)$ hysteresis loop and (**b**) primary and secondary scanning $\psi(s)$ loop predicted by Scott et al.'s [469] empirical hysteresis model

10.9 Modeling Hysteresis in the Retention Curve

Hysteresis in variably saturated porous media describes the dependence of the capillary pressure curve on the flow direction and history of wetting and drying. It can be caused by a number of mechanisms within the pore scale such as entrapment of air, shape of pore space (ink-bottle effect) and hysteresis in the contact angle [38, 473]. As the consequence the hysteretic behavior leads to a nonunique relationship between the pressure head ψ and the saturation s in the retention function $\psi(s)$ and $s(\psi)$, respectively (Fig. 10.4). During gravity drainage when the saturation and the pressure head are monotonically decreasing, the retention curve is still a unique function. However, when drying intermediately reverses into a wetting process and vice versa, $\psi(s)$ is no longer unique due to the hysteretic behavior. Similar effects can be observed during infiltration which exhibits a nonunique, but different $\psi(s)$ relationship, when wetting is reverses into a drying process and vice versa.

The unsaturated flow follows a *main wetting* and *main drying* retention curve $\psi^w(s)$ and $\psi^d(s)$ (or equivalent $s^w(\psi)$ and $s^d(\psi)$), respectively, when the porous medium is wetted from the residual saturation s_r or drained from the saturated state at maximum saturation s_s, respectively. Once a wetting or drying process is reversed while following the main hysteresis curve, the retention curve follows a *primary* hysteresis curve. Now, further reversals can occur, which leads to secondary and higher-order scanning curves (Fig. 10.4).

Various models for describing hysteresis in the retention curves have been developed [321, 411, 469, 554]. They can be grouped into physically based models and empirical models. Among them, empirical analytical models have shown most relevance in modeling practice due to their robustness and flexibility. In such a modeling approach it is assumed that the primary, secondary and higher-order scanning curves can be scaled from the main hysteresis curve. A very useful scaling approach has been introduced by Scott et al. [469], hereafter referred to as Scott et al.'s hysteresis model, which is based on analytic retention relations, such as the van Genuchten parametric model as described in Appendix D. Scott et al.'s scaling method [321, 469] can be further generalized and even applied to spline approximation retention curves [147].

We explain the Scott et al.'s hysteresis model along the analytic van Genuchten (VG) retention relations (D.2) and (D.5) of Appendix D, where its application to other analytic relationships becomes similar. For the VG curve it is

$$s_e = \frac{s - s_r}{s_s - s_r} = \begin{cases} \dfrac{1}{(1 + |\alpha\psi|^n)^m} & \text{for} \quad \psi < 0 \\ 1 & \text{for} \quad \psi \geq 0 \end{cases} \tag{10.102}$$

and

$$\psi = -\frac{1}{\alpha}\left(s_e^{-\frac{1}{m}} - 1\right)^{\frac{1}{n}} \quad \text{for} \quad 0 < s_e < 1 \tag{10.103}$$

We denote the main drying curve $s(\psi)$ or $\psi(s)$ by $s^d(\psi)$ and $\psi^d(s)$, respectively, and the main wetting curve by $s^w(\psi)$ and $\psi^w(s)$, respectively (Fig. 10.4). The main hysteresis loop for a VG retention is then described by the parameter vector $(s_s^w, s_r^w, \alpha^w, n^w, m^w)$ for the main wetting curve and $(s_s^d, s_r^d, \alpha^d, n^d, m^d)$ for the main drying curve, requiring in total ten curve parameters. If we restrict the approach to a *closed* hysteresis loop it can be assumed that the residual and maximum saturations of the main wetting curve are equal to those for the main drying curve such that

$$\begin{aligned} s_r^w &= s_r^d = s_r \\ s_s^w &= s_s^d = s_s \end{aligned} \tag{10.104}$$

We note that an extension to the case where $s_s^w \neq s_s^d$ is possible [321]. The assumption (10.104) reduces the total parameter set to eight parameters for describing the main hysteresis loop applied to the VG relation.

To compute the scanning curves denoted by $s^{w\star}(\psi)$ or $\psi^{w\star}(s)$ for wetting and $s^{d\star}(\psi)$ or $\psi^{d\star}(s)$ for drying, we employ Scott et al.'s hysteresis model, in which $s^{w\star}(\psi)$ is scaled from the main wetting curve $s^w(\psi)$ and $s^{d\star}(\psi)$ is scaled from the main drying curve $s^d(\psi)$. In doing so, drying scanning curves are obtained by using the VG parameter vector $(s_s^\star, s_r, \alpha^d, n^d, m^d)$ in (10.102), where $s_s^\star$ replaces s_s and has the effect of scaling the drying curve to pass through the reversal point indexed by Δ, giving (cf. Fig. 10.4b)

$$s_s^\star = \frac{s_\Delta - s_r\big(1 - s_e^d(\psi_\Delta)\big)}{s_e^d(\psi_\Delta)} \tag{10.105}$$

where s_Δ is the saturation at the reversal point and $s_e^d(\psi_\Delta)$ is the effective saturation (D.2) on the main drying curve at the reversal pressure head ψ_Δ, i.e., $s_e^d(\psi_\Delta)$ is described by (10.102) with $\psi = \psi_\Delta$ and parameters $(s_s, s_r, \alpha^d, n^d, m^d)$.

In a similar manner we can obtain any wetting scanning curve when the VG parameter vector $(s_s, s_r^\star, \alpha^w, n^w, m^w)$ is used, in which $s_r^\star$ is obtained from passing the main wetting curve through the reversal point, viz. (cf. Fig. 10.4b),

$$s_r^\star = \frac{s_\Delta - s_s s_e^w(\psi_\Delta)}{1 - s_e^w(\psi_\Delta)} \tag{10.106}$$

where $s_e^w(\psi_\Delta)$ is evaluated from (10.102) by using $\psi = \psi_\Delta$ and $(s_s, s_r, \alpha^w, n^w, m^w)$. A typical scanning hysteresis loop is illustrated in Fig. 10.4b, consisting of a primary wetting curve and secondary drying curve predicted by (10.106) and (10.105), respectively. In the VG parametric model all scanning curves have the form of (10.102) or (10.103). For other analytic retention relations such as described in Appendix D, Scott et al's hysteresis model can be analogously applied. To derive their corresponding scanning curves the specific parameter sets are modified by (10.105) and (10.106) for drying and wetting, respectively. Having defined the scanning curves of retention for drying and wetting, their first derivatives in form of moisture capacity and inverse moisture capacity can be obtained analogously to the expressions as given in Appendix D for the main curves.

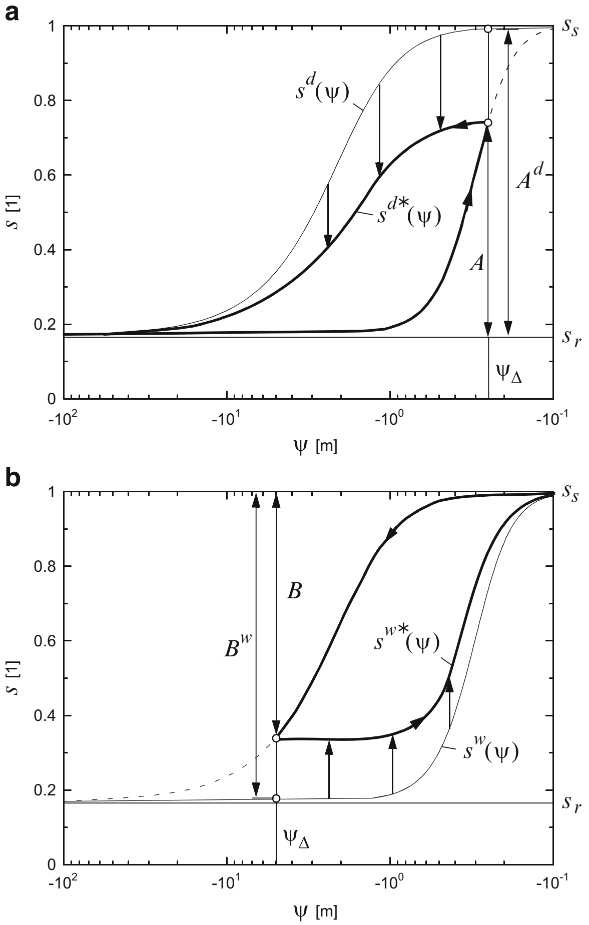
When using spline approximations for the retention curve as introduced in Sect. D.4 of Appendix D, saturation values s are obtained directly from $\psi-$values and effective saturation s_e is not involved. Here, a modification to Scott et al.'s hysteresis model is required as proposed in [147]. Assuming again a maximum saturation s_s common to both main curves and assuming an asymptotic minimum (residual) saturation s_r also common to both main curves, the pressure head ψ_Δ at a reversal point is used to define a linear scaling according to Fig. 10.5. We find for the reversal from wetting to drying (Fig. 10.5a)

$$\begin{aligned} s(\psi_\Delta) - s_r &= c^d\big(s^d(\psi_\Delta) - s_r\big) \\ c^d &= \frac{A}{A^d} = \frac{s(\psi_\Delta) - s_r}{s^d(\psi_\Delta) - s_r} \end{aligned} \tag{10.107}$$

and for the reversal from drying to wetting (Fig. 10.5b)

$$\begin{aligned} s_s - s(\psi_\Delta) &= c^w\big(s_s - s^w(\psi_\Delta)\big) \\ c^w &= \frac{B}{B^w} = \frac{s_s - s(\psi_\Delta)}{s_s - s^w(\psi_\Delta)} \end{aligned} \tag{10.108}$$

Fig. 10.5 Scanning spline-approximated $s(\psi)$ curve in a hysteretic loop for (**a**) drying curve $s^{d\star}(\psi)$ and (**b**) wetting curve $s^{w\star}(\psi)$



where c^d and c^w represent scaling factors. The required scanning curves are then defined by

$$s^{d\star}(\psi) = c^d s^d(\psi) + (1 - c^d)s_r \quad \text{for} \quad \psi < \psi_\Delta \tag{10.109}$$

for drying and

$$s^{w\star}(\psi) = c^w s^w(\psi) + (1 - c^w)s_s \quad \text{for} \quad \psi_\Delta < \psi < 0 \tag{10.110}$$

for wetting.

For the numerical implementation of the hysteresis models individual scanning curves are determined and recorded for each mesh node and time step. To each global node i and time plane n a hysteresis index $\kappa_{i,n}$ is assigned according to

$$\kappa_{i,n} = \begin{cases} +1 & \text{if node } i \text{ is wetting} \\ -1 & \text{if node } i \text{ is drying} \end{cases} \tag{10.111}$$

At node i and time plane n the flow direction is recorded as

$$\kappa_{i,n} = \text{sign}(\psi_{i,n} - \psi_{i,n-1}), \quad n = 1, 2, \ldots \tag{10.112}$$

As long as the flow direction does not change, i.e., $\kappa_{i,n} = \kappa_{i,n-1}$, the computation can proceed with the most recent scanning curve. Usually, at initial time t_0 the simulation starts from a corresponding main curve at a pre-set direction $\kappa_{i,0}$ and IC $\psi_{i,0}$. Once the flow direction changes, i.e., $\kappa_{i,n} \neq \kappa_{i,n-1}$, the value of change in the pressure head is accumulated and stored in the vector $\boldsymbol{\psi}_{\Sigma}$ for each node i as

$$\psi_{\Sigma,i} := \psi_{\Sigma,i} + \frac{\psi_{i,n} - \psi_{i,n-1}}{\kappa_{i,n}} \tag{10.113}$$

where $\boldsymbol{\psi}_{\Sigma}$ has been zeroed at the last reversal (or at initial state). Note that (10.113) accumulates only consecutive changes in the flow direction. Now, a (new) reversal occurs when the accumulated change $\psi_{\Sigma,i}$ at node i meets the following criterion

$$\psi_{\Sigma,i} > \epsilon_{\kappa} \tag{10.114}$$

where ϵ_{κ} is set to a dimensional small positive value (e.g., 10^{-3} m) to avoid oscillations in the flow directions. If (10.114) is satisfied the new reversal point is fixed for the node i with $\psi_{\Delta,i} = \psi_{i,n}$, the new scanning curve is determined based on the just detected reversal pressure head $\psi_{\Delta,i}$ and flow direction $\kappa_{i,n}$. The accumulated change $\psi_{\Sigma,i}$ is again reset to zero and the computation proceeds by using the new scanning curve.

Finally, it has to be noted that in opposite to the retention curves the relative permeability relations $k_r(s)$ usually exhibit only little or negligible hysteresis [321]. In particular, in the van Genuchten-Mualem (VGM) relationship (D.26) of Appendix D, where $m = 1 - 1/n$ is substituted, a hysteretic $k_r(s)$−dependency is implicitly given by the retention parameters and any hysteresis in $k_r(s)$ of the VGM parametric model is only due to differences between the pore size distribution parameters n^w and n^d for wetting and drying, respectively [339]. In the case that $n^w = n^d = n$ there is no more hysteresis in $k_r(s)$, however, hysteretic VG retention curves still arise if $\alpha^w \neq \alpha^d$.

10.10 Treatment of Prescribed Time-Varying Porosity

In many mining and underground construction cases, transient material properties play a significant role to mimic the effects of an excavation or infilling progress for porous-medium bodies. For example, an open pit is excavated below the water table

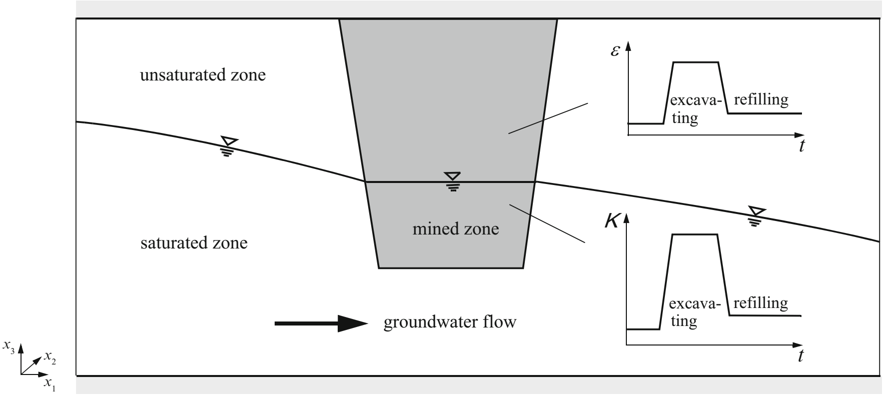


Fig. 10.6 Open pit excavation and refilling progress mimicked by prescribed time-varying porosity $\varepsilon(t)$ and hydraulic conductivity $\boldsymbol{K}(t)$ for the mined zone

and becomes later refilled or a long-wall coal mining involves backfilling of the mined zones with the excavation residues, which naturally changes the hydraulic properties in form of time-varying, *a priori* known relationships for porosity $\varepsilon = \varepsilon(t)$ and hydraulic conductivity $\boldsymbol{K} = \boldsymbol{K}(t)$ in these zones [458], Fig. 10.6. While a transient behavior of any material property can be easily implemented via standard time series without further specific physical considerations, there is the only exception for a prescribed porosity function $\varepsilon(t)$, which is subjected to the time derivation in the general storage term $\partial(\varepsilon s \rho)/\partial t$, cf. (3.225) with (3.219). As a consequence, the storage term has to be developed different to the standard formulations given above by (3.242) and (3.246). Unlikely, we have to use

$$\frac{\partial(\varepsilon s \rho)}{\partial t} = \varepsilon s \frac{\partial \rho}{\partial t} + \rho \varepsilon \frac{\partial s}{\partial t} + \rho s \frac{\partial \varepsilon}{\partial t} = \varepsilon s \rho_0^2 \gamma g \frac{\partial h}{\partial t} + \rho \varepsilon \frac{\partial s}{\partial t} + \rho s \frac{\partial \varepsilon}{\partial t} \qquad (10.115)$$

where the liquid compressibility γ is introduced according to Sect. 3.10.4, while the porosity term remains untreated and will not be expressed in a term of skeleton compressibility as usually done by invoking (3.246). As a result, we obtain an alternative Richards' equation written in the $h-s-$form as

$$s\, S_o^\star \frac{\partial h}{\partial t} + \varepsilon \frac{\partial s}{\partial t} - \nabla \cdot \left[k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e}) \right] = Q_h + Q_{hw} + Q_{\mathrm{EOB}} - s\dot{\varepsilon} \qquad (10.116)$$

which is somewhat different to the original formulation (10.5) in respect to (1) the modified specific storage coefficient $S_o^\star = \rho_0 g \varepsilon \gamma \leq S_o$ encompassing only liquid compressibility effects and (2) the explicit appearance of the derivative porosity term $s\dot{\varepsilon}$ in the RHS to be expressed via a prescribed time series, viz.,

$$\varepsilon = f(t), \quad \dot{\varepsilon} = \tfrac{\partial f(t)}{\partial t} \qquad (10.117)$$

where $f(t)$ corresponds to a continuous (once-differentiable) input function to be specified for excavation or infilling measures. If $\dot{\varepsilon} > 0$ the excavation progresses in time as a sink which leads to a decrease of the hydraulic head in the zone. On the other hand, if an infilling progresses with $\dot{\varepsilon} < 0$ the derivative porosity term acts as an additional source associated with an increase of the hydraulic head in the corresponding zone.

The computation of the alternative Richards' equation (10.116) can be analogously performed by using the numerical techniques as described in the preceding sections. The only difference is in the usually reduced specific storage coefficient $S_o^\star \leq S_o$ and in the additional sink term $-s\dot{\varepsilon}$ of a prescribed porosity history. For the latter an extra vector $\boldsymbol{F}^\star$ will appear in the resulting spatio-temporally discretized equation on the RHS in the form

$$\begin{aligned} \boldsymbol{F} &:= \boldsymbol{F} + \boldsymbol{F}^\star \\ \boldsymbol{F}^\star \;&=\; F_i^\star = -\sum_e \int_{\Omega^e} N_i s^e \dot{\varepsilon^e} d\Omega^e \end{aligned} \tag{10.118}$$

where $\boldsymbol{F}$ is the usual RHS vector defined in (10.32) or (10.61). In application to the above time stepping schemes the temporal evaluation of the extra term $\boldsymbol{F}^\star$ is executed at the current time plane $n+1$ as

$$\boldsymbol{F}^\star_{n+1} = F^\star_{i,n+1} = -\sum_e \int_{\Omega^e} N_i \sum_l (N_l^e s_{l,n+1}^e) \tfrac{\partial f(t)}{\partial t}\Big|_{n+1} d\Omega^e \tag{10.119}$$

Note that the extension in the RHS vector $\boldsymbol{F}$ by $\boldsymbol{F}^\star$ also enters in the Jacobian for the Newton iteration method, cf. Sect. 10.7.3.

10.11 Computation of Darcy Velocities and Flow Budget Analysis

The Darcy velocity and boundary flux computation for flow in variably saturated porous media follow the same principles as described in Sect. 9.7 for saturated flow. The only difference is in using the specific formulations in dependence on the chosen equations types. Thus, the discrete evaluation of Darcy velocities is performed in the $h-$ and $\psi-$formulation, respectively, as (cf. Sect. 8.19.1)

$$\begin{aligned} \boldsymbol{q}(\boldsymbol{x}, t_{n+1}) &= -\sum_j k_r(s_{n+1}) \boldsymbol{K} f_\mu \cdot \big[\nabla N_j(\boldsymbol{x})\, h_j(t_{n+1}) + \chi \boldsymbol{e}\big] \\ \boldsymbol{q}(\boldsymbol{x}, t_{n+1}) &= -\sum_j k_r(s_{n+1}) \boldsymbol{K} f_\mu \cdot \big[\nabla N_j(\boldsymbol{x})\, \psi_j(t_{n+1}) + (1+\chi) \boldsymbol{e}\big] \end{aligned} \tag{10.120}$$

where $h_j(t_{n+1}) = \boldsymbol{h}_{n+1}$ and $\psi_j(t_{n+1}) = \boldsymbol{\psi}_{n+1}$ are the known nodal hydraulic head and pressure head, respectively, and $k_r(s_{n+1})$ is evaluated by the known discrete saturation at nodal points $s_{n+1} = \sum_j N_j(\boldsymbol{x})s_j(t_{n+1})$ at time plane $n+1$. Smoothing techniques as thoroughly described in Sect. 8.19.1 are used to derive continuous Darcy velocities at the nodal points.

To obtain a precise flow budget evaluation the CBFM, as introduced in Sect. 8.19.2, is applied to the specific weak formulations of the Richards' equation. While the evaluation for the standard $h-$based form (10.47) can be done analogously to (9.64), giving now

$$\begin{aligned}\int_\Gamma N_i\, q_n\, d\Gamma = -\int_\Omega N_i(sS_o + \varepsilon C)\frac{\partial h}{\partial t} d\Omega - \\ \int_\Omega \nabla N_i \cdot [k_r(s)\boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega + \\ \int_\Omega N_i(Q_h + Q_{hw} + Q_{\mathrm{EOB}}) d\Omega \end{aligned} \tag{10.121}$$

the boundary fluxes for the mixed $h-s-$form (and the equivalent $\psi - s-$form) of the Richards' equation must be evaluated from the original weak statement (10.25) written as

$$\begin{aligned}\int_\Gamma N_i\, q_n\, d\Gamma = -\int_\Omega N_i\, sS_o\frac{\partial h}{\partial t} d\Omega - \int_\Omega N_i \varepsilon\frac{\partial s}{\partial t} d\Omega - \\ \int_\Omega \nabla N_i \cdot [k_r(s)\boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega + \\ \int_\Omega N_i(Q_h + Q_{hw} + Q_{\mathrm{EOB}}) d\Omega \end{aligned} \tag{10.122}$$

where $h = \sum_j N_j\, h_j$ and $s = \sum_j N_j\, s_j$ are known at t_{n+1}. Finally, matrix systems result to solve the consistent boundary flux vector $\boldsymbol{q}_n$ for the specific forms from

$$\boldsymbol{M} \cdot \boldsymbol{q}_n = \begin{cases} -\boldsymbol{O}^\dagger \cdot \dot{\boldsymbol{h}} - \boldsymbol{D}^\dagger \cdot \boldsymbol{h} + \boldsymbol{F}^\dagger & h\text{−form} \\ -\boldsymbol{O} \cdot \dot{\boldsymbol{h}} - \boldsymbol{B} \cdot \dot{\boldsymbol{s}} - \boldsymbol{D}^\dagger \cdot \boldsymbol{h} + \boldsymbol{F}^\dagger & h-s\text{−form} \end{cases} \tag{10.123}$$

where

$$\begin{aligned} \boldsymbol{M} &= M_{ij} &&= \int_\Gamma N_i N_j d\Gamma \\ \boldsymbol{O} &= O_{ij} &&= \delta_{ij}\int_\Omega sS_o N_i d\Omega \\ \boldsymbol{O}^\dagger &= O_{ij}^\dagger &&= \delta_{ij}\int_\Omega (sS_o + \varepsilon C) N_i d\Omega \\ \boldsymbol{B} &= B_{ij} &&= \delta_{ij}\int_\Omega \varepsilon N_i d\Omega \\ \boldsymbol{D}^\dagger &= D_{ij}^\dagger &&= \int_\Omega \nabla N_i \cdot (k_r(s)\boldsymbol{K} f_\mu \cdot \nabla N_j) d\Omega \\ \boldsymbol{F}^\dagger &= F_i^\dagger &&= \int_\Omega N_i(Q_h + Q_{\mathrm{EOB}}) d\Omega - \int_\Omega \nabla N_i \cdot (k_r(s)\boldsymbol{K} f_\mu \cdot \chi \boldsymbol{e}) d\Omega - \\ & && \quad Q_w(t)\Big|_i \end{aligned} \tag{10.124}$$

in which $\dot{\boldsymbol{h}}$, $\dot{\boldsymbol{s}}$ and $\boldsymbol{h}$ are known at the evaluation time t_{n+1}. In a flow budget analysis the integral boundary balance flux $\boldsymbol{Q}_n$ is directly evaluated at each boundary node by

$$\boldsymbol{Q}_n = -\boldsymbol{M} \cdot \boldsymbol{q}_n = \begin{cases} \boldsymbol{O}^\dagger \cdot \dot{\boldsymbol{h}} + \boldsymbol{D}^\dagger \cdot \boldsymbol{h} - \boldsymbol{F}^\dagger & h\text{−form} \\ \boldsymbol{O} \cdot \dot{\boldsymbol{h}} + \boldsymbol{B} \cdot \dot{\boldsymbol{s}} + \boldsymbol{D}^\dagger \cdot \boldsymbol{h} - \boldsymbol{F}^\dagger & h - s\text{−form} \end{cases} \tag{10.125}$$

10.12 Upstream Weighting

Forsyth and Kropinski [166] pointed out the necessity of upstream weighting in unsaturated-saturated flow problems to avoid spurious local maxima and minima at coarse mesh sizes. Monotonicity considerations were applied to find appropriate evaluation points for the relative permeability 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 permeability 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 ADE as discussed in Sect. 8.14.

Different approaches exist in unsaturated flow modeling for the representation of material properties. Forsyth and Kropinski [166], Šimůnek et al. [553] or Oldenburg and Pruess [397] 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 permeability k_r can be directly located between adjacent nodes. Such schemes have proved to be unconditionally monotone [166].

The present upstream weighting method is based on an elemental representation of material properties. We use the following procedure to find appropriate upstream weighting points at an element level. A central weighting is equivalent to the influence coefficient method using a linear combination of nodal parameters at element level according to

$$k_r^e(\boldsymbol{\eta}, t) \approx \sum_J N_J^e(\boldsymbol{\eta}) k_{r,J}^e(t) \tag{10.126}$$

where $k_{r,J}^e$ is the relative permeability determined at local node J of element e and the element basis functions $N_J^e(\boldsymbol{\eta}) = N_J^e(\xi, \eta, \zeta)$ are evaluated at the element centroid ($\xi = \eta = \zeta = 0$). Instead of using the central position, we select an upstream position $\tilde{\boldsymbol{\eta}} = (\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ for computing the relative permeability via (10.126). The evaluation point $\tilde{\boldsymbol{\eta}}$ is used for Gauss integration in the matrix terms related to k_r and is similar to the Gauss-point-based upwind technique proposed by Hughes [266]. To determine the upstream local coordinates $\tilde{\boldsymbol{\eta}}$ in 2D and 3D elements the following method is applied.

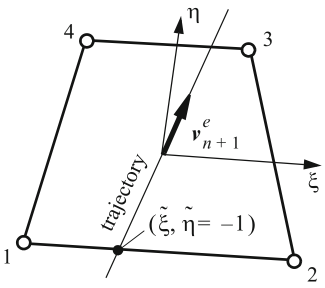


Fig. 10.7 Upstream local coordinates $(\tilde{\xi}, \tilde{\eta})$ in a 2D finite element

Based on the predicted pressure head ψ^p_{n+1} a specific flux can be computed at a central position of an element e, viz.,

$$\boldsymbol{v}^e_{n+1} = -\sum_J \nabla N^e_J(0,0,0)\Big(\psi^p_{J,n+1} + (1+\chi)e_J\Big) \tag{10.127}$$

and, the trajectory of the vector $\boldsymbol{v}^e_{n+1}$ can be easily found. Along the trajectory, in the upstream direction, the upstream position $\tilde{\boldsymbol{\eta}}$ is set at the intersection with the element border (Fig. 10.7). For the element e the relative permeability is now evaluated at the upstream point as

$$k^e_r(\boldsymbol{\eta}, t) = \sum_J N^e_J(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta}) k^e_{r,J}(t) \tag{10.128}$$

With the upstream point $(\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$ the relative permeability k^e_r is evaluated only along element edges. For instance, considering the situation in Fig. 10.7 for a 2D isoparametric finite element, $\tilde{\eta}$ is -1 and k^e_r, from (10.128), becomes independent of nodes 3 and 4, viz., $k^e_r = \frac{1}{2}[(1-\tilde{\xi})k^e_{r,1} + (1+\tilde{\xi})k^e_{r,2}]$.

10.13 Examples

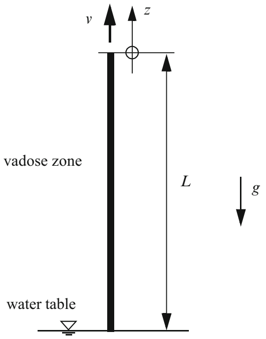
10.13.1 Variably Saturated Flow in a Homogeneous Soil Column

10.13.1.1 Gardner's Problem

An analytical solution for the steady-state pressure head distribution ψ above the water table in a 1D soil column (Fig. 10.8) has been presented by Gardner [185] in a form[7]:

[7] Two interesting results can be detected from (10.129):

Fig. 10.8 Sketch of the unsaturated solution domain above the water table



$$\psi(z) = -\tfrac{1}{\alpha}\ln\Big(\tfrac{1}{K}[(K+v)e^{-\alpha(L+z)} - v]\Big) \tag{10.129}$$

where z is the vertical coordinate (positive upward), v is the evaporation (exfiltration rate) positively directed along z (note that v corresponds to an infiltration rate if negative), K is the (scalar) saturated hydraulic conductivity, α is the sorptive number by assuming an exponential relationship for the relative permeability $k_r = \exp(\alpha\psi)$ [417] according to (D.39) (with $\psi_a = 0$) of Appendix D, L is the height of the column, and ln() represents the natural logarithm.

The analytical solution (10.129) will be compared with numerical results by using the parameters as summarized in Table 10.4. The numerical simulations are based on the standard $h-$form of the Richards' equation since the problem is steady-state and no specific requirements for a precise storage term approximation arise. Linear finite elements are used. The comparisons for the pressure head $\psi(z)$ are shown in Table 10.5 for the cases of constant evaporation $+v$ and constant infiltration $-v$. As revealed the agreement is quite perfect.

1. We can ask which flux is concerned to force the pressure head zero everywhere? It can be easily shown from (10.129) that such a situation occurs if the infiltration has the amount of the saturated conductivity, i.e., $v = -K$
2. We also can ask which flux is concerned to make the pressure head ψ infinity at the soil surface $z = 0$, i.e., $\psi(0) = \infty$? This should occur for a certain rate v which represents the theoretically maximum evaporative flux $v_{\max}$. The pressure head ψ becomes infinity at $z = 0$ if the argument of the logarithm of (10.129) goes to zero. It implies that

$$\frac{v_{\max}}{K} = e^{-\alpha L}\Big(\frac{v_{\max}}{K} + 1\Big)$$

and leads to a solution of the theoretically *maximum evaporative flux* as

$$v_{\max} = \frac{K}{e^{\alpha L} - 1}.$$

Table 10.4 Parameters and conditions used for Gardner's problem

Quantity	Symbol	Value	Unit
Column length	L	1	m
Saturated conductivity	K	10^{-7}	$\mathrm{m\,s^{-1}}$
Exponential parametric model[a] (D.16), (D.39)			
Sorptive number	α	1	$\mathrm{m^{-1}}$
BC's			
Neumann-type BC at top	$q_h = v$	$\pm 8.64 \cdot 10^{-4}$	$\mathrm{m\,d^{-1}}$
Dirichlet-type BC at bottom	ψ_D	0	m
FEM			
Space increment	Δz	10^{-2}	m
RMS error tolerance	ϵ	10^{-3}	1

[a] Note that $\psi_a = 0$

Table 10.5 Comparison of pressure head ψ for the cases of a constant evaporation and infiltration

Elevation above water table $(L+z)$ (m)	Evaporation $+v$		Infiltration $-v$	
	Analytical (m)	Numerical (m)	Analytical (m)	Numerical (m)
0.95	−1.122654	−1.122649	−0.802813	−0.802814
0.85	−0.993830	−0.993826	−0.724280	−0.724281
0.75	−0.868446	−0.868443	−0.644110	−0.644111
0.65	−0.746020	−0.746018	−0.562398	−0.562399
0.55	−0.626153	−0.626151	−0.479238	−0.479240
0.45	−0.508510	−0.508509	−0.394725	−0.394726
0.35	−0.392810	−0.392809	−0.308948	−0.308948
0.25	−0.278814	−0.278813	−0.221993	−0.221994
0.15	−0.166316	−0.166316	−0.133946	−0.133946
0.05	−0.055140	−0.055140	−0.044886	−0.044886

10.13.1.2 Celia et al.'s Problem

Celia et al. [72] introduced a transient unsaturated flow problem to benchmark modeling approaches for a strong infiltration front development in a homogeneous soil column. Celia et al. [72] used the modified Picard method for the mixed $\psi - s-$based form of the Richards' equation as described in Sect. 10.5.3. They discretized the column of 1 m length by using the spatial increments for a dense and a coarse grid with $\Delta z = 0.5\,\mathrm{cm}$ and $\Delta z = 2.5\,\mathrm{cm}$, respectively. In [72] dense-grid simulations were performed with a constant time increment $\Delta t = 60\,\mathrm{s}$, which means that their 'best' solutions for a simulation time of 1 day were obtained after 1,440 time steps plus a number of unreported Picard steps. The used parameters and conditions for the benchmark problem are summarized in Table 10.6.

We use the same spatial discretization with linear finite elements as applied by Celia et al. [72], however, prefer the FE/BE or AB/TR predictor-corrector

Table 10.6 Parameters and conditions used for Celia et al.'s problem [72, 470]

Quantity	Symbol	Value	Unit
Column length	L	1	m
Saturated conductivity	K	$9.22 \cdot 10^{-5}$	$\mathrm{m\,s^{-1}}$
Porosity	ε	0.368	1
Specific storage coefficient	S_o	0	$\mathrm{m^{-1}}$
Maximum saturation	s_s	1	1
Residual saturation	s_r	0.277	1
van Genuchten-Mualem (VGM) parametric model[a]	(D.3), (D.26)		
Pore size distribution index	n	2	1
Fitting coefficient	α	3.35	$\mathrm{m^{-1}}$
IC and BC's			
Initial condition (IC)	ψ_0	-10	m
Dirichlet-type BC at top	ψ_D^T	-0.75	m
Dirichlet-type BC at bottom	ψ_D^B	-10	m
FEM			
Space increment (fine and coarse grid)	Δz	0.5 and 2.5	cm
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE and AB/TR)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^\star$	10^{-4}	$\mathrm{m^3\,d^{-1}}$
Simulation time period	t_{end}	1	d

[a] Note that it is generally used: $m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

time stepping strategies based on either the PVST with Newton method or the Picard scheme for the mixed $\psi - s$–form of the Richards' equation. A comparison with Celia et al.'s results is shown in Fig. 10.9a and reveals very good agreements (note that Celia et al.'s results are picked from a table presented in [470], where only selected sample points are listed). In Fig. 10.9b the dense and coarse grid solutions are compared to illustrate spatial discretization effects. As shown, a significant phase lead and a somewhat smeared ψ–profile result.

The time behavior of the residual error $\|\boldsymbol{R}_{n+1}\|_{\mathrm{RMS}}$ and the total balance error TBE(t) (10.87) are plotted in Fig. 10.10 for the PVST and FE/BE scheme. At $t_{\mathrm{end}} = 1\,\mathrm{d}$ the simulation terminates with $\mathrm{TBE}(t_{\mathrm{end}}) \approx 10^{-6}\,\mathrm{m}^3$ by using a RMS error tolerance of $\epsilon = 10^{-4}$ and residual error tolerance of $\epsilon_2^\star = 10^{-4}\,\mathrm{m^3\,d^{-1}}$ (10.86). These PVST simulations with variably switched primary variable for ψ or s need about 400 time steps and 450 Newton iterations in total. In contrast, the Picard method applied to the mixed $\psi - s$–formulation with the pressure head ψ as the general primary variable embedded in the FE/BE and AB/TR predictor-corrector scheme requires 1,350 and 1,000 time steps, respectively, and about 1,500 Picard iterations in total for both time stepping schemes, which is more than thrice the computational effort of the PVST computations. This represents the time step demand also reported by Celia et al. [72] who used 1,440 steps. The same order of time and iteration steps occurs when the solution is performed with the standard

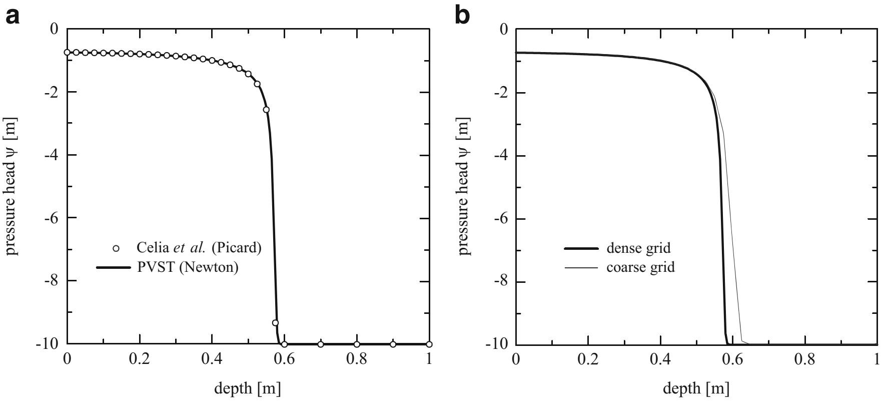


Fig. 10.9 Pressure head ψ profiles at $t = 1$ day: (**a**) dense grid solution obtained for PVST (both FE/BE and AB/TR scheme) in comparison with Celia et al.'s results [72, 470] and (**b**) dense versus coarse grid solution computed by PVST

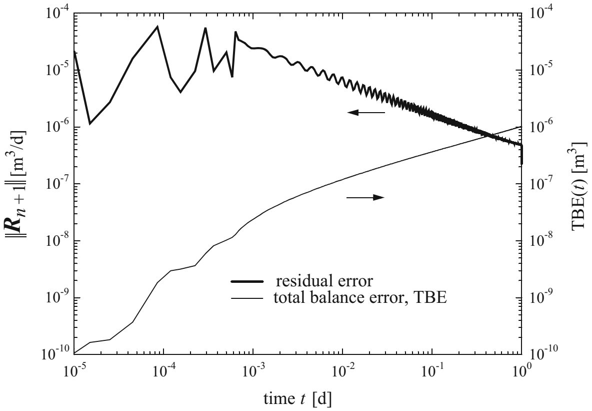


Fig. 10.10 History of residual error $\|\boldsymbol{R}_{n+1}\|_{\text{RMS}}$ and total balance error TBE(t) occurred for PVST and FE/BE scheme by using error tolerances $\epsilon = 10^{-4}$ and $\epsilon_2^\star = 10^{-4}$ m^3 d^{-1}

$h-$form of the Richards' equation. Using the dense grid all formulations give the results of Fig. 10.9a in a good agreement. In a conclusion, however, it is obvious that PVST is the most efficient solution strategy. More investigations and comparison to an empirical target-based time stepping strategy can be found in [141].

10.13.1.3 Williams et al.'s Problem

Williams et al. [566] analyzed various numerical approaches for simulating sharp front dynamics in 1D soil columns. Their introduced test cases possess high nonlinearity in the governing unsaturated-saturated flow equations, which makes

Table 10.7 Parameters and conditions used for Williams et al.'s problem [566]

Quantity	Symbol	Value	Unit
Column length	L	10	m
Saturated conductivity	K	$5.833 \cdot 10^{-5}$	m s^{-1}
Porosity	ε	0.301	1
Specific storage coefficient	S_o	10^{-6}	m^{-1}
Maximum saturation	s_s	1	1
Residual saturation	s_r	0.30897	1
van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26)			
Pore size distribution index	n	4.264	1
Fitting coefficient	α	5.47	m^{-1}
IC and BC's			
Initial condition (IC)	h_0	0	m
Dirichlet-type BC at top	h_D^T	10.1	m
Dirichlet-type BC at bottom	h_D^B	0	m
FEM			
Space increment	Δz	1.25	cm
Initial time step size	Δt_0	10^{-10}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^\star$	10^{-4}	$\mathrm{m}^3\,\mathrm{d}^{-1}$
Simulation time period	t_{end}	0.18	d

[a] Note that it is generally used: $m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

the problems difficult to solve via standard methods. We choose one representative example from Williams et al.'s benchmarks [566], which exhibits a sharp front development under both unsaturated and saturated conditions. The height of the homogeneous soil column is 10 m. The used parameters and conditions for this benchmark problem are summarized in Table 10.7.

For the present simulations the column is discretized by 800 linear quadrilateral elements. Time-constant Dirichlet-type BC's for the hydraulic head h on top and bottom of the column are imposed. We simulate this problem by PVST, the Celia et al.'s method for the mixed form of Richards' equations with Picard iteration and the standard form of the Richards' equation with Picard iteration and chord slope approximation. The present computations use always the automatic FE/BE predictor-corrector time stepping. A residual error control is applied to PVST and Celia et al.'s method. The results in form of pressure head profiles in time obtained via PVST are compared in Fig. 10.11 with the findings by Williams et al. [566] presenting both dense-grid and coarse-grid solutions. It is evident that the present PVST results agree almost perfectly with Williams et al.'s dense-grid solution. The PVST simulation takes 2,693 variable times steps and results a total balance error TBE($t = 0.18$ d) (10.87) of $\mathcal{O}(10^{-6})$ m^3 by using a RMS error tolerance of $\epsilon = 10^{-4}$ and residual error tolerance of $\epsilon_2^\star = 10^{-4}$ $\mathrm{m}^3\,\mathrm{d}^{-1}$ (10.86). The Celia et al.'s method leads to comparable pressure head profiles, however, requires a double time step number (actually, 5,129 variable time steps) to achieve

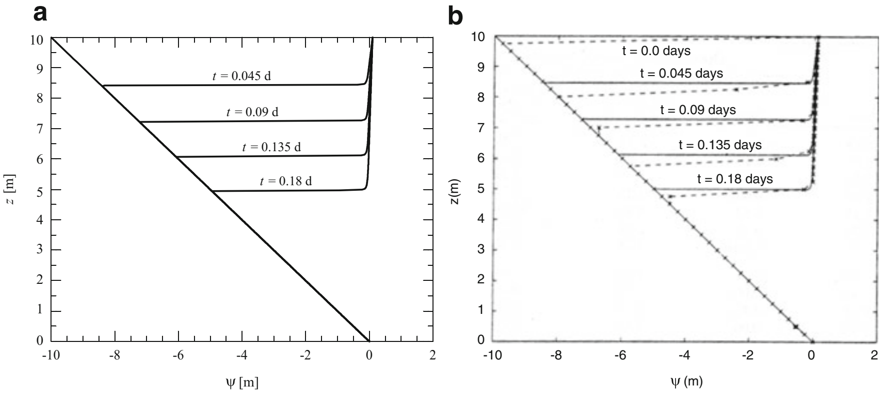


Fig. 10.11 Pressure head ψ profiles at different times t in days: (**a**) present solutions obtained by PVST with automatic FE/BE predictor-corrector time stepping and (**b**) Williams et al.'s results [566] obtained for two meshes (*solid lines* – dense grid with 25,601 nodes and fixed time step size of $\Delta t = 1.56 \cdot 10^{-6}$ d, *dashed lines* – coarse grid with only 41 nodes and fixed time step size of $2 \cdot 10^{-4}$ d)

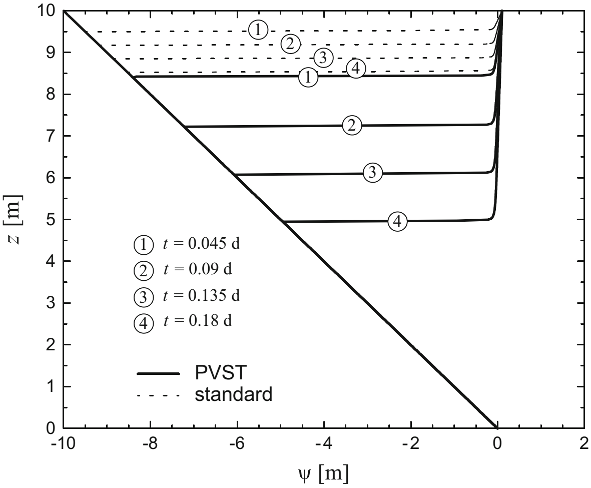


Fig. 10.12 Pressure head ψ profiles at different times t in days: comparison between PVST solution and results obtained by the standard form of Richards' equation, both simulated on the same mesh and by using automatic FE/BE predictor-corrector time stepping

an equivalent total balance error of $\mathcal{O}(10^{-6})$ m^3. Contrarily, the standard form of the Richards' equation leads to awfully bad results, revealing a significant mass balance inaccuracy as evidenced in Fig. 10.12. The extreme lag in the pressure profiles indicates a considerable loss of mass for this type of formulation, notwithstanding the taken higher number of time steps (actually, 6,202 variable time steps). We find

Table 10.8 Parameters and conditions used for Lenard et al.'s problem [85, 340]

Quantity	Symbol	Value	Unit
Column length	L	0.72	m
Saturated conductivity	K	$3.3056 \cdot 10^{-4}$	$\mathrm{m\,s^{-1}}$
Porosity	ε	0.36	1
Specific storage coefficient	S_o	10^{-4}	$\mathrm{m^{-1}}$
Maximum saturation	$s_s = s_s^d = s_s^w$	1	1
Residual saturation	$s_r = s_r^d = s_r^w$	0.17	1
van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26)			
Wetting curve			
Pore size distribution index	n^w	5.25	1
Fitting coefficient	α^w	8.4	$\mathrm{m^{-1}}$
Drying curve			
Pore size distribution index	n^d	5.25	1
Fitting coefficient	α^d	4.2	$\mathrm{m^{-1}}$
IC and BC			
Initial condition (IC) of hydraulic head (with $z = 0$ at bottom)	h_0	0.695	m
Dirichlet-type BC at bottom (time-dependent hydraulic head)	$h_D(t)$	Table 10.9	m
FEM			
Space increment	Δz	3.6	mm
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^\star$	10^{-4}	$\mathrm{m^3\,d^{-1}}$
Simulation time period	t_{end}	10	h

[a] Note that it is generally used: $m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

that the standard form is unsuitable for simulating such type of sharp fronts with rapid changes in saturations over short distances and time scales.

10.13.1.4 Lenhard et al.'s Problem: Hysteretic Flow

Lenhard et al. [340] conducted an air-water flow experiment in a 72-cm vertical soil column, where the water table fluctuates. This column experiment was recomputed by Clausnitzer [85] in using a 1D FEM. We will compare the present finite element solutions with Clausnitzer's results for both hysteretic and nonhysteretic conditions. For a discussion of the simulation results with the experimental findings we refer to [85]. The used material parameters and conditions are listed in Table 10.8.

The column is filled by a homogeneous sandy material with a saturated hydraulic conductivity of $1.19\,\mathrm{m\,h^{-1}}$ and porosity of 0.36. Initially, most of the column is fully saturated, with the water table positioned at 0.695 m (25 mm below the top of the column). At the bottom of the soil column the hydraulic head h is varied in time. Its boundary curve is listed in Table 10.9. Results will be compared for five

Table 10.9 Time-dependent hydraulic head $h_D(t)$ at the bottom of the soil column

Time t (min)	$h_D(t)^a$ (m)
0	0.695
125	0.070
175	0.070
245	0.420
295	0.420
345	0.170
395	0.170
505	0.720
600	0.720

[a] Changes between the time stages are linear

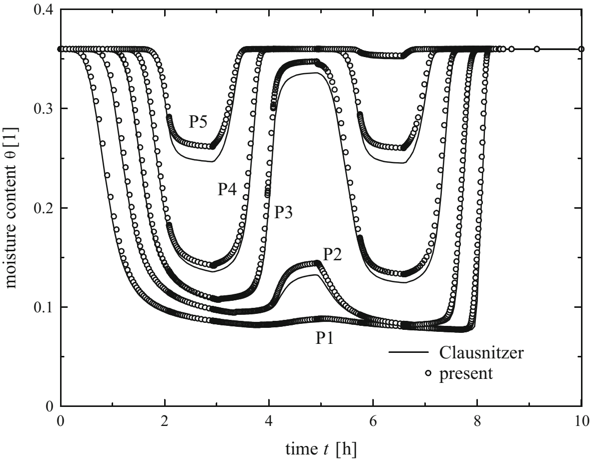


Fig. 10.13 History of moisture content θ simulated at different vertical locations P1 through P5 in the column for hysteretic conditions: comparison between present and Clausnitzer's results [85]

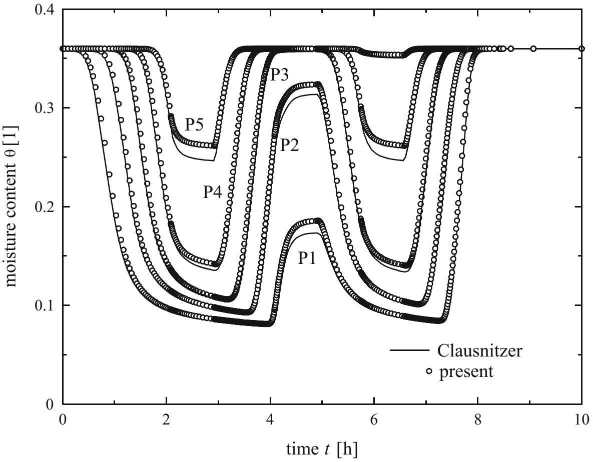


Fig. 10.14 History of moisture content θ simulated at different vertical locations P1 through P5 in the column for nonhysteretic conditions: comparison between present and Clausnitzer's results [85]

observation points located at (P1) 0.69 m, (P2) 0.59 m, (P3) 0.49 m, (P4) 0.39 m, and (P5) 0.29 m, measured from the soil bottom. For the present simulations both the standard and the mixed form of the Richards' equation with the automatic FE/BE predictor-corrector scheme and Picard iteration strategy are used. A comparison of the FEFLOW results with the solutions obtained by Clausnitzer [85] gives a good agreement as revealed in Fig. 10.13. It is to be noted that Clausnitzer used a coarser mesh with a width of $\Delta z = 1\,\mathrm{cm}$ and a fully implicit scheme with variable time steps based on the standard form of the Richards' equation. The hysteresis effect on the moisture content $\theta(t)$ for the present problem can be seen in comparing the solutions for the hysteretic case (Fig. 10.13) with the solutions for the nonhysteretic case (Fig. 10.14). The results for nonhysteretic conditions are obtained by using the parameters of the drying main curve. Note that the *moisture content*, or soil water content, is defined as the product of saturation s and porosity ε:

$$\theta = s\varepsilon \tag{10.130}$$

where the maximum moisture content is $\theta_s = s_s\varepsilon$ and the residual moisture content is $\theta_r = s_r\varepsilon$.

10.13.2 Variably Saturated Flow in an Inhomogeneous Soil Column

10.13.2.1 Van Genuchten's Problem

Van Genuchten [470, 537, 538] describes results for moisture movement in a layered soil. A soil column with a length of 170 cm includes four 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 (Fig. 10.15). Tables 10.10 and 10.11 summarize the parameters and conditions used in the computations. A time-varying Neumann-type flux BC is imposed on the surface with constant $q_h = -0.25\,\mathrm{m\,d^{-1}}$ at $t \leq 1\,\mathrm{d}$ (infiltration) and constant $q_h = +0.005\,\mathrm{m\,d^{-1}}$ at $t > 1\,\mathrm{d}$ (exfiltration). On the bottom a gradient-type BC of $q_h^{\nabla} = -1(-1) \cdot K_{\text{bottom}} = 4\,\mathrm{m\,d^{-1}}$ is imposed to allow a free (gravitationally driven) drainage of the soil column.

A comparison between the present solutions and the results obtained by van Genuchten [537] who used a Hermitian-finite element approach is exhibited in Figs. 10.16 and 10.17. Figure 10.16 displays the simulated moisture-content profiles during the infiltration period at $t \leq 1$ d. The moisture-content histories during the redistribution phase at $t > 1\,\mathrm{d}$ in the soil column are compared in Fig. 10.17. As shown the agreement of the results is nearly perfect.

Since the IC of $\psi_0 = -3.5\,\mathrm{m}$ does not imply a very dry soil, the original van Genuchten's problem is not particularly difficult to solve and all formulations

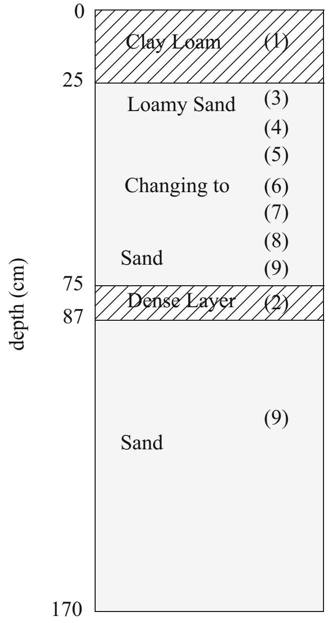


Fig. 10.15 Layered soil profile of van Genuchten's problem [470, 537]

Table 10.10 Parameters and conditions used for van Genuchten's problem [470, 537]

Quantity	Symbol	Value	Unit
Column length	L	170	cm
Column layered by soils (Fig. 10.15) listed in Table 10.11			
Clay loam (0–25 cm)	Soil type 1		
Dense layer (75–87 cm)	Soil type 2		
Loamy sand (25–75 cm)	Soil gradually changing from type 3 to 9		
Sand (87–170 cm)	Soil type 9		
IC and BC's			
Initial condition (IC)	ψ_0	-3.5	m
Neumann-type BC at surface (infiltration/exfiltration)	q_h	$\begin{cases} -0.25 & \text{at } t \leq 1\,\text{d} \\ +0.005 & \text{at } t > 1\,\text{d} \end{cases}$	m d^{-1}
Gradient-type BC at bottom (free drainage)	$q_h^{\nabla} = K$	4	m d^{-1}
FEM			
Space increment	Δz	1	cm
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE and AB/TR)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^{\star}$	10^{-4}	m^3 d^{-1}
Simulation time period	t_{end}	8	d

and schemes are successful. However, 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)

Table 10.11 Material properties of soils by using the van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26) applied to van Genuchten's problem [470, 537]

Soil type	Quantity	Depth (cm)	ε (1)	s_r (1)	α (m^{-1})	n (1)	K ($10^{-4} m\,s^{-1}$)	S_o (m^{-1})
1	Clay loam	0–25	0.5400	0.3704	0.800	1.8	0.029	$4 \cdot 10^{-5}$
2	Dense layer	75–87	0.4000	0.6250	0.900	3.0	0.012	$5 \cdot 10^{-6}$
3	Loamy sand	25–32	0.4700	0.3617	1.000	2.0	0.087	$1 \cdot 10^{-5}$
4		32–41	0.4611	0.3494	1.306	2.178	0.154	$1 \cdot 10^{-5}$
5		41–50.5	0.4500	0.3333	1.080	2.4	0.237	$1 \cdot 10^{-5}$
6		50.5–59	0.4400	0.3182	1.120	2.6	0.313	$1 \cdot 10^{-5}$
7		59–66	0.4311	0.3041	1.156	2.778	0.379	$1 \cdot 10^{-5}$
8		66–71	0.4244	0.2931	1.182	2.911	0.430	$1 \cdot 10^{-5}$
9	Sand	$\begin{cases} 71-75 \\ 87-170 \end{cases}$	0.4200	0.2857	1.200	3.0	0.463	$1 \cdot 10^{-5}$

[a] Note that it is generally used: $s_s = 1, m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

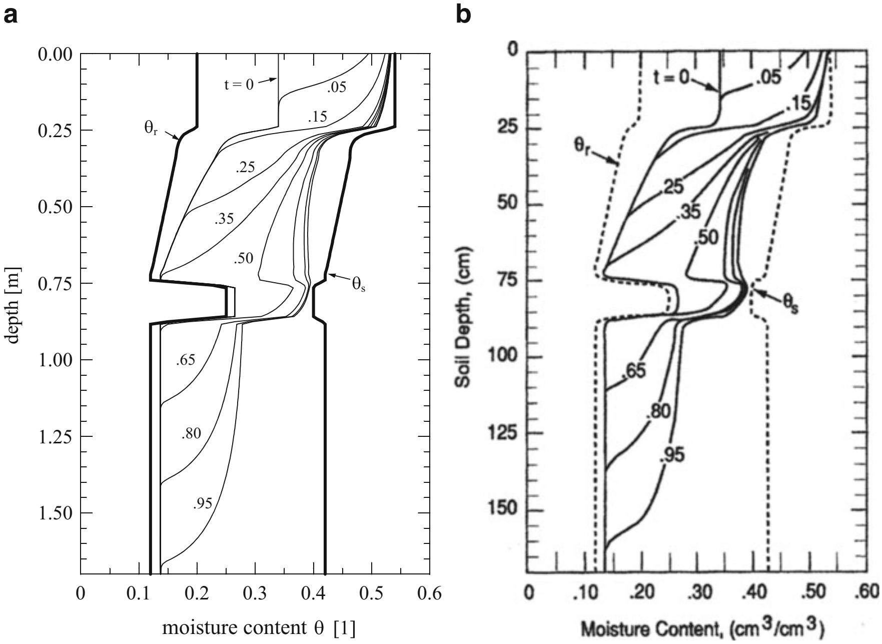


Fig. 10.16 Simulated moisture-content ($\theta = s\varepsilon$) profiles during infiltration: (**a**) present solutions and (**b**) van Genuchten's results [470, 537], time in days

under low and extremely low initial pressure heads ψ_0. Using the PVST with the FE/BE predictor-corrector scheme the computed saturation profiles at $t = 1$ d are shown in Fig. 10.18a for different ψ_0. As expected, at very dry IC's the saturation profile remains unchanged, proving thus the good conservative properties of the

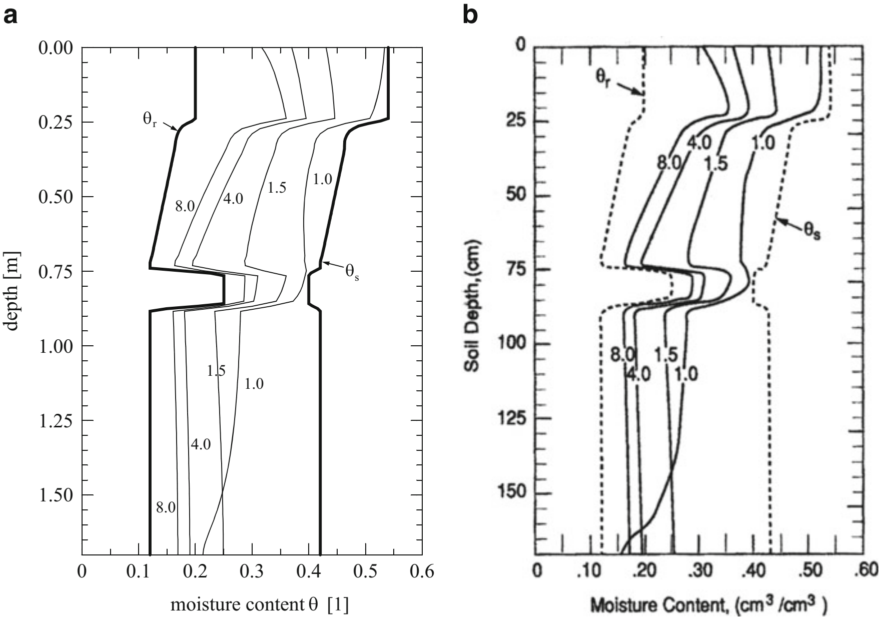


Fig. 10.17 Simulated moisture-content ($\theta = s\varepsilon$) profiles during redistribution: (**a**) present solutions and (**b**) van Genuchten's results [470, 537], time in days

PVST. Practically any arbitrary large value of ψ_0 can be enforced. In contrast to this, standard formulations using the pressure head ψ or hydraulic head h 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 10.18b shows the results for both the mixed $\psi - s-$form with Newton iteration and the standard $h-$form with Picard iteration and chord slope approximation. As seen at low initial pressure head ($\psi_0 = -3.5\,\text{m}$) the schemes yield the same results. However, already for $\psi_0 = -10\,\text{m}$ the standard $h-$form reveals mass-conservative problems (phase lag). The phase lag error dramatically grows at lower initial pressure heads as evidenced in Fig. 10.18b for $\psi_0 = -10^3\,\text{m}$. On the other hand, the conservative mixed $\psi - s-$form provides better results, though not without a phase lag error at $\psi_0 = -10^3\,\text{m}$ (Fig. 10.18b) in comparison to the good PVST results (Fig. 10.18a). We were not able to find convergent solutions for both the mixed $\psi - s-$form and the standard $h-$form at lower pressure head values ($\psi_0 < -10^3$ m).

Table 10.12 summarizes the solution effort in terms of time steps and required number of total iterations for different schemes depending on the initial pressure head ψ_0. The PVST is successful for all ψ_0 considered, while the schemes using the pressure head ψ or hydraulic head h as primary variable (mixed Newton $\psi - s-$form and standard Picard $h-$form) have shown unsuitable for very dry conditions

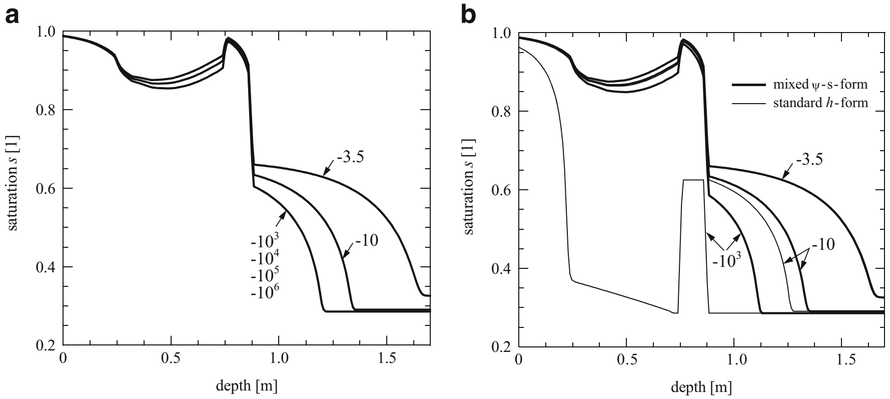


Fig. 10.18 Saturation distribution at $t = 1\,\mathrm{d}$ simulated from various initial pressure heads ψ_0 measured in (m) by using: (**a**) PVST with FE/BE predictor scheme and (**b**) Newton mixed $\psi - s$—form and the standard Picard iteration h—form, both with FE/BE predictor scheme

Table 10.12 Solution effort for different schemes (simulation time 1 day, FE/BE predictor corrector with RMS error tolerance $\epsilon = 10^{-4}$ and maximum rate of time step change of $\Xi = 2$)

			Primary variable ψ or h			
	PVST		Mixed $\psi - s$—form, Newton		Standard h—form, Picard	
	Total time	Total Newton	Total time	Total Newton	Total time	Total Picard
ψ_0 (m)	Steps	Steps[a]	Steps	Steps[a]	Steps	Steps[a]
-3.5	358	360	634	638	643	648
-10	676	684	1,824	2,112	1,760	2,021
-10^3	1,510	2,187	4,202	4,792	1,128	1,472
-10^4	1,990	3,254	Failed		Failed	
-10^5	2,180	3,858	Failed		Failed	
-10^6	2,696	4,988	Failed		Failed	

[a] Including rejected steps

$\psi_0 < -10^3$ m. The PVST is always superior under very dry conditions [141]. Since the predictor-corrector method is controlled by the temporal discretization error, 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. In the simulations the total balance error TBE($t = 1$ d) (10.87) has found of $\mathcal{O}(10^{-6})$ m^3 by using a RMS error tolerance of $\epsilon = 10^{-4}$ and residual error tolerance of $\epsilon_2^\star = 10^{-4}$ m^3 d^{-1} (10.86).

Table 10.13 Material properties of soils by using the exponential parametric model[a] (D.15), (D.38) applied to Brunone et al.'s two-layer problem [61]

Layer	Thickness (cm)	ε (1)	s_r (1)	α (m^{-1})	K (m s^{-1})	S_o (m^{-1})
1	20	0.4	0.15	10	$2.778 \cdot 10^{-6}$	10^{-4}
2	80	0.4	0.15	10	$2.778 \cdot 10^{-5}$	10^{-4}

[a] Note that it is generally used: $s_s = 1$ and $\psi_a = 0$

Table 10.14 Parameters and conditions used for Brunone et al.'s two-layer problem

Quantity	Symbol	Value	Unit
Column length	L	100	cm
Column layered by two soils as quantified in Table 10.13			
IC and BC's[a]			
Initial condition (IC)	ψ_0	Variable[b]	m
Neumann-type BC at surface	q_h	-0.228	m d^{-1}
Dirichlet-type BC at bottom	h_D	-2	m
FEM			
Space increment	Δz	2.5	mm
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^\star$	10^{-4}	m^3 d^{-1}
Simulation time period	t_{end}	40	h

[a] Origin of z—coordinate is at the surface directed upward

[b] Obtained at steady state with infiltration of $4.54 \cdot 10^{-4}$ cm h^{-1}

10.13.2.2 Brunone et al.'s Two-Layer Problem with Srivastava and Yeh's Analytical Solution

Brunone et al. [61] studied a vertical infiltration into a 1 m-deep soil profile consisting of two layers with parameters listed in Table 10.13. The IC is a steady-state pressure head distribution corresponding to a constant infiltration of $4.54 \cdot 10^{-4}$ cm h^{-1}, while the surface flux at $t = 0$ is abruptly changed to an infiltration rate of 0.95 cm h^{-1}. The lower BC is subjected to a pressure head ψ of -100 cm. Imposed boundary flux at the soil surface corresponds to a rain intensity of 0.95 cm h^{-1}. The nodal spacing of the computational mesh is $\Delta z = 2.5$ mm. The used simulation parameters and conditions are summarized in Table 10.14.

The present two-layer problem can be solved analytically by using the method proposed by Srivastava and Yeh [490]. Profiles of pressure head ψ and moisture content θ at selected times during the infiltration process as computed by Srivastava and Yeh's analytical method are shown in Fig. 10.19a. The analytical results are compared with the numerical findings shown in Fig. 10.19b. It reveals a very good agreement with the analytical results. For the numerical simulation we use the PVST with automatic FE/BE predictor-corrector time integration. It takes only 153 time steps with 155 Newton steps in total.

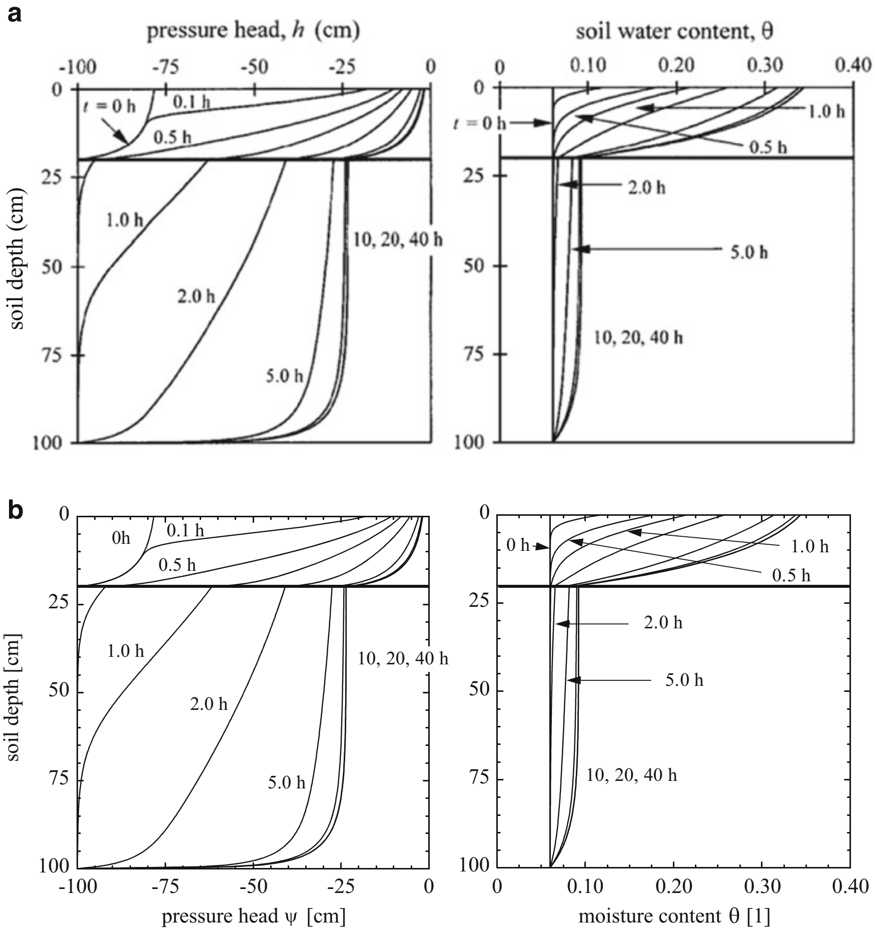


Fig. 10.19 Profiles of pressure head ψ (*left*) and moisture content θ (*right*) at selected times t (h): (**a**) Srivastava and Yeh's analytical solution given by Brunone et al. [61] and (**b**) present numerical solutions achieved with PVST and automatic FE/BE predictor-corrector time integration. The *horizontal bold line* indicates the location of the interface between the two soil layers

10.13.2.3 Matthews et al.'s Problem: Two-Layer Soil Contrast

The effect of soil contrast was considered by Matthews et al. [358] for the vertical infiltration into a 40-cm-deep soil profile consisting of two 20 cm thick layers of fine and coarse soils. They used a Newton iteration method for solving the s−form of Richards' equation. In their study two test cases are simulated: test case 1 comprises the fine over coarse soil and test case 2 comprises the coarse over fine soil. The parameters for these two soils listed in Table 10.15 are highly contrasting.

Imposed boundary flux at the soil surface corresponds to a rain intensity of 2 cm d^{-1}. At the bottom boundary a free drainage BC is imposed. The IC is given

Table 10.15 Material properties of soils by using the van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26) applied to Matthews et al.'s two-layer problem [358]

Soil type	Quantity	Thickness (cm)	ε (1)	s_r (1)	α (m^{-1})	n (1)	K ($\mathrm{m\,s}^{-1}$)	S_o (m^{-1})
Coarse	Berino fine sand	20	0.3658	0.0782	2.80	2.2390	$6.2616 \cdot 10^{-5}$	10^{-4}
Fine	Glendale clay	20	0.4686	0.2262	1.04	1.3954	$1.5162 \cdot 10^{-6}$	10^{-4}

[a] Note that it is generally used: $s_s = 1, m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

Table 10.16 Parameters and conditions used for Matthews et al.'s problem: test case 1 and test case 2

Quantity	Symbol	Value	Unit
Column length	L	40	cm
Column layered by two soils as quantified in Table 10.15:			
** Test case 1 – fine soil over coarse soil*			
** Test case 2 – coarse soil over fine soil*			
IC and BC's			
Initial condition (IC)	ψ_0	-10	m
Neumann-type BC at surface	q_h	-0.02	$\mathrm{m\,d}^{-1}$
Gradient-type BC at bottom (free drainage)	$q_h^{\nabla} = K$	$\begin{cases} 0.131 & \text{(test case 1)} \\ 5.410 & \text{(test case 2)} \end{cases}$	$\mathrm{m\,d}^{-1}$
FEM			
Space increment	Δz	1	mm
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^{\star}$	10^{-4}	$\mathrm{m}^3\,\mathrm{d}^{-1}$
Simulation time period	t_{end}	3	d

by a constant pressure head ψ of -10 m. For the present simulation a mesh with constant $\Delta z = 1$ mm is used. The applied simulation parameters and conditions are summarized in Table 10.16.

Figure 10.20 compares the present FEFLOW simulations with the numerical results given by Matthews et al. [358] in form of the moisture content θ distribution over depth for test case 1 (fine soil over coarse soil) and test case 2 (coarse soil over fine soil). As seen in both cases the agreement is rather well. The FEFLOW simulations are based on PVST by using automatic FE/BE predictor-corrector time integration. They required about 440 Newton steps for test case 1 and 340 Newton steps for test case 2 in total. At final time t_{end} of 3 days the solutions terminate with a total balance error (10.87) of $\mathrm{TBE}(t_{\mathrm{end}}) \approx 10^{-7}\,\mathrm{m}^3$ by using a RMS error tolerance of $\epsilon = 10^{-4}$ and residual error tolerance of $\epsilon_2^{\star} = 10^{-4}\,\mathrm{m}^3\,\mathrm{d}^{-1}$ (10.86). The sharpness of the wetting profiles at the soil interface is evident (Fig. 10.20). Water reached the soil interface at about 1 day in both cases. For the test case 1 (fine soil over coarse soil) we observe that water does not penetrate the underlying coarse soil layer to a depth of 5 cm until $t = 1.5$ d, which can be recognized as a capillary barrier effect, where water is held at the interface by capillary forces (see Sect. 10.13.5 for a further discussion of capillary barrier modeling).

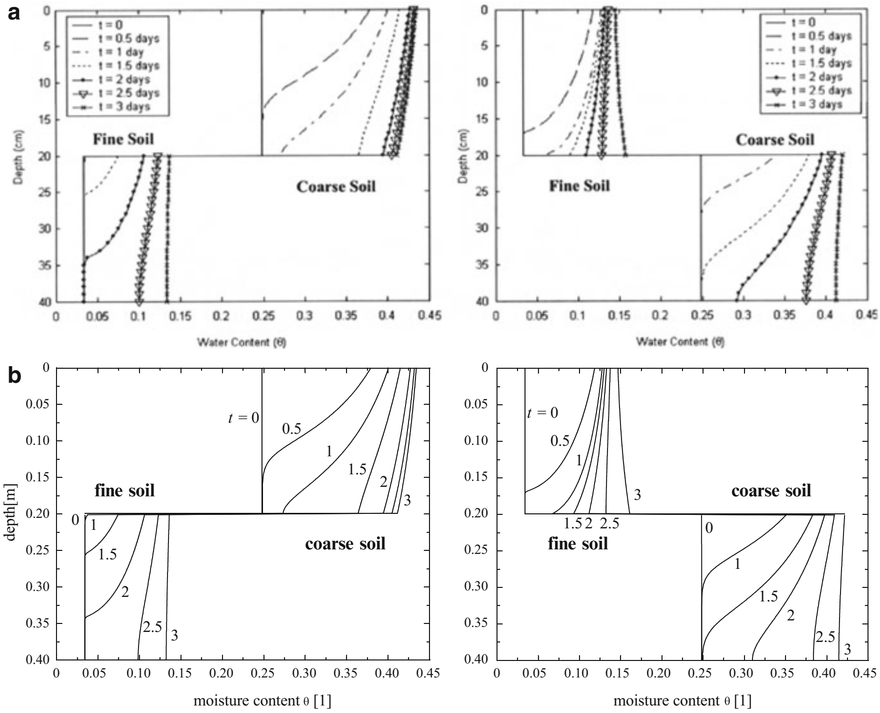


Fig. 10.20 Computed profiles of moisture content θ at selected times t (d) for test case 1 (*left*) and test case 2 (*right*): (**a**) Matthews et al.'s results [358] by using a Newton scheme in solving the s—form of Richards' equation and (**b**) present numerical solutions achieved with PVST and automatic FE/BE predictor-corrector time integration

10.13.3 Forsyth and Kropinski's Problem: Infiltration in a Large Caisson

The infiltration process in a large caisson consisting of heterogeneous materials at dry IC's has been thoroughly studied by Forsyth et al. [167] and in a modified version by Forsyth and Kropinski [166]. This model problem was used by Diersch and Perrochet [141] as a benchmark test example to compare the predictor-corrector-based PVST approach with Forsyth et al.'s target-based time-stepping PVST strategy. Figure 10.21 presents a schematic view of the 2D cross-sectional model problem. All boundaries are impervious except for the infiltration boundary section on top. Table 10.17 lists the material properties used for the different zones of the domain and the conditions applied in the simulations for the Forsyth and Kropinski's problem [166].

For the present simulations PVST with FE/BE predictor-corrector time stepping is preferred. We start at first with the same coarse spatial discretization of 90 × 21 quadrilateral elements (1,890 nodes) as in Forsyth and Kropinski's

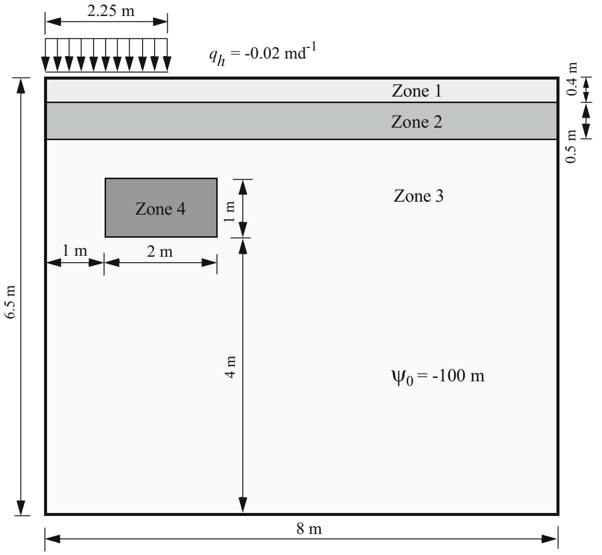


Fig. 10.21 Model problem of infiltration in a large caisson (Modified from [167])

Table 10.17 Parameters and conditions used for Forsyth and Kropinski's problem [166, 167]

Quantity	Symbol	Value	Unit
Caisson measure (length; height)		8; 6.5	m
Specific storage coefficient	S_o	10^{-4}	m^{-1}
Zones of porous materials (Fig. 10.21) listed in Table 10.18			
IC and BC			
Initial condition (IC)	ψ_0	-100	m
Neumann-type BC at surface	q_h	-0.02	$m\,d^{-1}$
FEM			
2D meshes of quadrilateral elements in various resolutions			
Initial time step size	Δt_0	10^{-3}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^\star$	10^{-4}	$m^3 d^{-1}$
Simulation time period	t_{end}	30	d

simulations [166]. A comparison of the simulated saturation profiles is shown in Fig. 10.22 indicating mesh effects. Although using the same mesh, differences at material interfaces and at the bottom of the caisson are detected. These obviously result from different nodal spacing at these locations [141]. The present PVST and FE/BE adaptive time stepping procedure required 1,202 time steps with 2,015 Newton steps in total.

Upstream weighting (Sect. 10.12) can be used to damp out the spurious oscillations in the saturation distributions. Figure 10.23 compares the present upstream solution with Forsyth and Kropinski's result. The agreement is quite good. Both

Table 10.18 Material properties by using the van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26) applied to Forsyth and Kropinski's problem [166, 167]

Zone[b]	ε (1)	s_r (1)	α (m^{-1})	n (1)	K (m s^{-1})
1	0.368	0.2771	3.34	1.982	$9.153 \cdot 10^{-5}$
2	0.351	0.2806	3.63	1.632	$5.445 \cdot 10^{-5}$
3	0.325	0.2643	3.45	5	$4.805 \cdot 10^{-5}$
4	0.325	0.2643	4.45	5	$4.805 \cdot 10^{-4}$

[a] Note that it is generally used: $s_s = 1$, $m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

[b] Displayed in Fig. 10.21

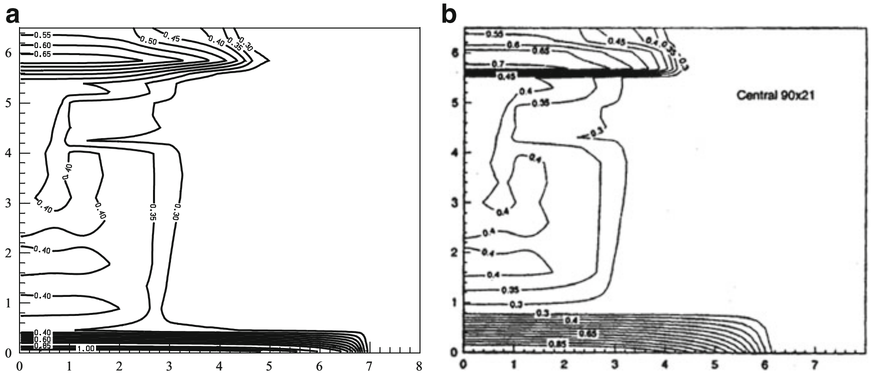


Fig. 10.22 Saturation contours at $t = 30$ d simulated with 90 × 21 nodal meshing: (**a**) present results and (**b**) Forsyth and Kropinski's results [166], lengths in (m)

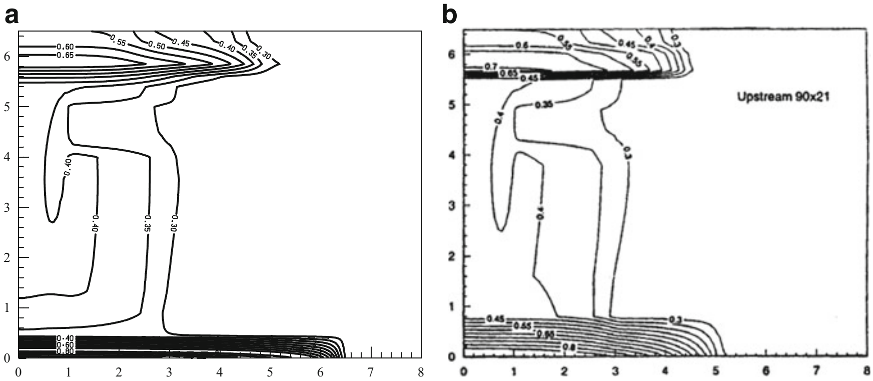


Fig. 10.23 Saturation contours at $t = 30$ d simulated with 90 × 21 nodal meshing for upstream weighting: (**a**) present results and (**b**) Forsyth and Kropinski's results [166], lengths in (m)

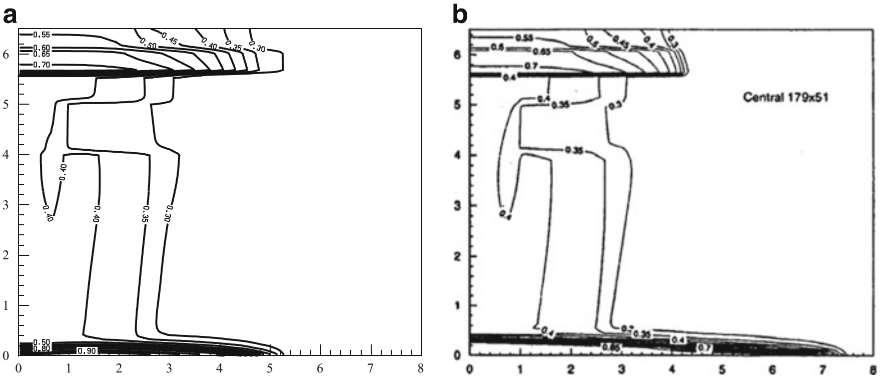


Fig. 10.24 Saturation contours at $t = 30\,\text{d}$ for more appropriate mesh resolutions: (**a**) present results simulated with 21×90 nodal meshing and (**b**) Forsyth and Kropinski's results [166] obtained with 179 × 51 nodal meshing, lengths in (m)

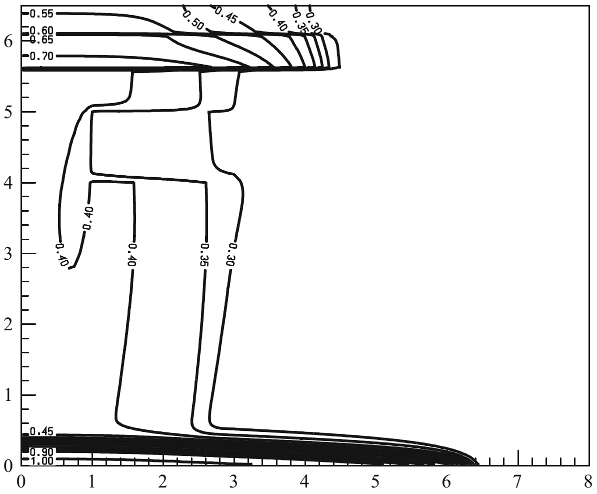


Fig. 10.25 Saturation contours at $t = 30\,\text{d}$ for the dense triangular mesh (28,917 nodes), lengths in (m)

upstream techniques damp out the wiggles appearing in the standard (central) weighting solutions (Fig. 10.22). Differences in the lag of the saturation profile are probably due to the different nodal spacing used in the present and Forsyth and Kropinski's solutions.

A more appropriate meshing of the problem (i.e., 21 × 90 instead of 90 × 21) can considerably improve the results as evidenced in Fig. 10.24. The solution can be compared to the results obtained with a dense triangular mesh (56,960 triangles with 28,917 nodes) shown in Fig. 10.25. This dense mesh is generated by splitting each quadrilateral into two triangles followed by a double total refinement into four triangles (20 × 89 × 2 × 4 × 4). Sharper saturation contours occur at the

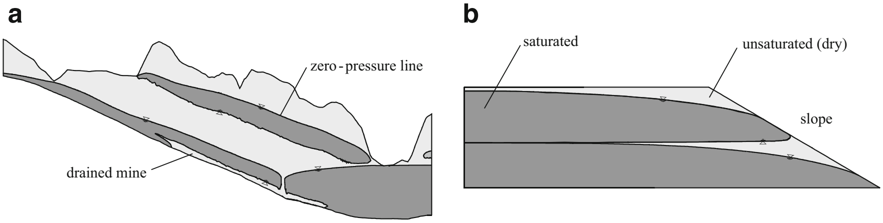


Fig. 10.26 Typical perched water situations, cross-sectional view of aquifer-aquitard-aquifer systems: (**a**) drained underground mine and (**b**) seepage at a slope

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. 10.24, more discussions are given in [141]. Remarkably, Forsyth and Kropinski predict a lead in the saturation pattern (Fig. 10.24b). The present FEFLOW results have been confirmed by Aricò et al. [11] in using a different numerical approach. In checking the total mass balance errors TBE(t), (10.87), we estimate TBE($t = 30$ d) of $\mathcal{O}(10^{-6})$ m^3 for the present simulations.

10.13.4 Perched Water Table Problems

10.13.4.1 Aquifer-Aquitard-Aquifer Test Case

Perched water situations can occur in many practical cases (Fig. 10.26) which require the application of unsaturated-saturated modeling approaches. Often in regional flow modeling the vertical spatial resolution is aligned to the stratigraphic units and we have to ask whether such a meshing is able to model a perched water situation whenever potentially achievable. To examine perched water computations let us start with a simplified 1D paradigm of an aquifer-aquitard-aquifer system as described in Fig. 10.27 for which an analytical solution exists in steady-state [139]. The used parameters are listed in Table 10.19.

Perched water occurs if the infiltrating flux $|q_h|$ is larger than the smallest gravitational efflux between layers, i.e.,

$$K_2 > |q_h| > K_1 \tag{10.131}$$

Due to continuity the vertical flux is $q(z) = \text{const}$. If the water becomes perched in the upper aquifer, saturated conditions must exist in the aquitard and the Darcy law reads

$$q_h = -K_1 \frac{h_1}{d_1} \tag{10.132}$$

where h_1 is the hydraulic head measured in the aquitard. Perched water in the upper aquifer rises up to a height of $h_w = z_w$. Thus, we obtain with (10.132)

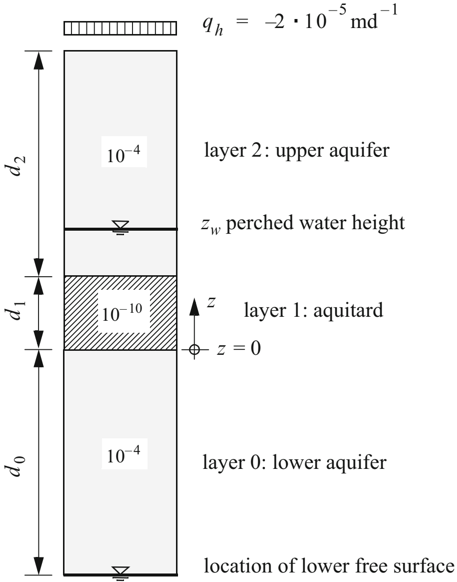


Fig. 10.27 Paradigm of a 1D aquifer-aquitard-aquifer system for determining the perched water height z_w above the aquitard at a given stationary infiltration rate $q_h = -2 \cdot 10^{-5}\,\mathrm{m\,d^{-1}}$

Table 10.19 Parameters used for the 3-layer (aquifer-aquitard-aquifer) paradigm

Layer	No.	Thickness d (m)	Conductivity K (m s^{-1})
Lower aquifer	0	$d_0 = 50$	$K_0 = 10^{-4}$
Aquitard	1	$d_1 = 10$	$K_1 = 10^{-10}$
Upper aquifer	2	$d_2 = 50$	$K_2 = 10^{-4}$

$$q_h = -K_2 \frac{h_w - h_1}{z_w - z_1} = -K_2 \frac{z_w + \frac{d_1}{K_1} q_h}{z_w - z_1} \tag{10.133}$$

From (10.133) we find the formula for the perched water height z_w as

$$z_w = \frac{q_h d_1 \left(\frac{1}{K_2} - \frac{1}{K_1}\right)}{1 + \frac{q_h}{K_2}} \tag{10.134}$$

For the case $K_2/K_1 \gg 1$ the perched water height z_w is approximately

$$z_w = -\frac{d_1}{K_1} q_h \tag{10.135}$$

Using the parameters for the paradigm of Table 10.19 we determine the exact perched water height at $z_w^{\text{exact}} = 23.148$ m.

Table 10.20 Test cases and computational results for the 3-layer (aquifer-aquitard-aquifer) paradigm

Case	Description	Parameters[a]	Height z_w (m)	Error δ [b] (%)
1	VGM model[c]	$\alpha = 4.1\,\mathrm{m}^{-1}$ $n = 2$	23.01	0.6
2	Linear model	$-\psi_c = 1\,\mathrm{m}$	22.39	3.3
3		$-\psi_c = 10\,\mathrm{m}$	18.56	19.8
4		$-\psi_c = 50\,\mathrm{m}$	Nonexistent	–

[a] It is generally used: $s_s = 1$, $s_r = 0.1$ and $\psi_a = 0$

[b] $\delta = |z_w - z_w^{\text{exact}}|/z_w^{\text{exact}} \cdot 100$; $z_w^{\text{exact}} = 23.148\,\mathrm{m}$

[c] The VGM model uses: $m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

Numerical experiments are performed by using the VGM parametric model, (D.3), (D.26), in comparison with the linear parametric model, (D.19), (D.42), typically used in classic groundwater modeling with fixed meshes in a form (cf. Sect. 9.5.4):

$$s_e = 1 - \frac{\psi}{\psi_c} \quad \text{and} \quad k_r = s_e \tag{10.136}$$

where $s_e = (s - s_r)/(1 - s_r)$ is the effective saturation and $-\psi_c$ is the capillary fringe thickness. The spatial discretization in regional models often follows stratigraphic units resulting in a coarse vertical resolution. In essence, this corresponds to a capillary fringe thickness $-\psi_c$ in the order of the vertical element size h^e, viz.,

$$-\psi_c = h^e \tag{10.137}$$

In this case, the transition in the saturation is smeared over the entire vertical element size, and accordingly, ψ_c becomes element-dependent and is no longer a 'physical' parameter. The consequences will be shown in the numerical experiments listed in Table 10.20. While test case 1 represents the 'physically correct' reference solution, test case 4 corresponds to an example of a large capillary fringe $-\psi_c$ typical for an approach, where the vertical mesh resolution is on the order of the stratigraphic units. The computed perched water height z_w for the four test cases are compared against the exact solution in Table 10.20. The resulting saturation profiles are displayed in Fig. 10.28. For case 1 a vertical discretization of $\Delta z = 0.1\,\mathrm{m}$ was applied. (Note, an equivalent linear model would need $-\psi_c \approx \frac{1}{\alpha}$ for a corresponding accuracy, where the discretization should require $\Delta z \leq -\psi_c = \frac{1}{\alpha}$).

It becomes evident that the error for the perched water height significantly increases with increasing $|\psi_c|$ (or equivalently with the increasing mesh coarseness $\Delta z \approx |\psi_c|$). In case 4 the upper aquifer (layer 2) is simulated to be completely unsaturated, resulting in a non-perched water situation. It indicates that the mesh must be sufficiently resolved in the vertical direction, clearly more than the

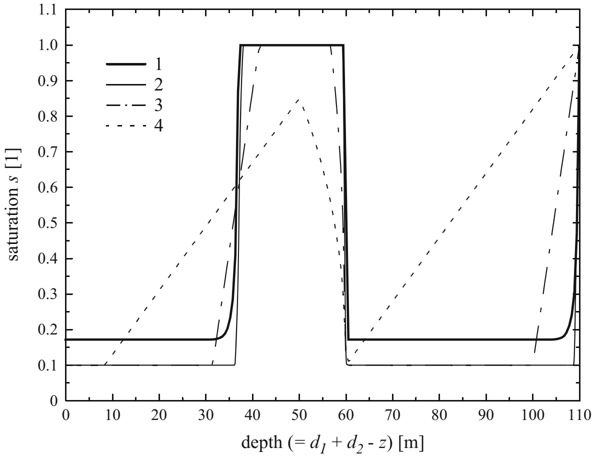


Fig. 10.28 Computed saturation profiles for the cases 1–4 of Table 10.20

stratigraphic units, otherwise the model is usually unable to predict a perched water table.

10.13.4.2 Kirkland et al.'s Problem

Kirkland et al. [311] presented a 2D problem of a developing perched water table surrounded by very dry unsaturated conditions. The problem is described in Fig. 10.29. Water infiltrates with a very large rate into an initially dry soil at $\psi_0 = -500\,\text{m}$ 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 used parameters and conditions of the problem are summarized in Table 10.21. The symmetric half of the domain is discretized in a 50×60 quadrilateral mesh (3,111 nodes) according to the spatial discretization used by Kirkland et al. [311] having spatial increments of $\Delta x = \Delta y = 5\,\text{cm}$. The present simulations are based on the PVST.

A comparison of the pressure head contours at 1 day with Kirkland et al.'s results reveals an acceptable agreement as displayed in Fig. 10.30. The present profile exhibits a slightly higher sharpness. While the zero pressure head contours agree quite well, the $-400\,\text{m}$ isoline of Kirkland et al.'s result is slightly ahead, forming a more diffusive vertical pressure front compared to the present solution. As further discussed in [141] the observed differences can mainly be attributed to temporal discretization effects. Typically, a smaller step number generates a phase lead and a smoother front. In the present simulations the PVST with the automatic FE/BE predictor-corrector time stepping took 1,211 time steps with 1,556 Newton steps in total. The TBE($t = 1$ d) balance error (10.87) was found to be of $\mathcal{O}(10^{-4})$ m^3.

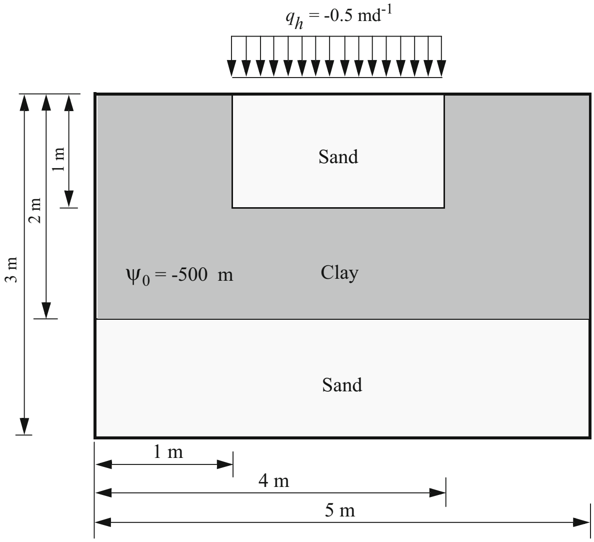


Fig. 10.29 Perched water table problem (Modified from [311])

Table 10.21 Parameters and conditions used for Kirkland et al.'s problem [311]

Quantity	Symbol	Value	Unit
Box measure (length; height)		5; 3	m
Specific storage coefficient	S_o	10^{-4}	m^{-1}
Properties of material zones (Fig. 10.29) listed in Table 10.22			
IC and BC			
Initial condition (IC)	ψ_0	-500	m
Neumann-type BC at surface	q_h	-0.5	m d^{-1}
FEM			
2D mesh of 50 × 60 quadrilateral elements for symmetric half			
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Residual error tolerance	$\epsilon_2^{\star}$	10^{-4}	$\text{m}^3\,\text{d}^{-1}$
Simulation time period	t_{end}	1	d

Table 10.22 Material properties by using the van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26) applied to Kirkland et al.'s problem [311]

Material[b]	ε (1)	s_r (1)	α (m^{-1})	n (1)	K (m s^{-1})
Sand	0.3658	0.07818	2.80	2.2390	$6.262 \cdot 10^{-5}$
Clay	0.4686	0.2262	1.04	1.3954	$1.516 \cdot 10^{-6}$

[a] Note that it is generally used: $s_s = 1, m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

[b] Zones of materials are displayed in Fig. 10.29

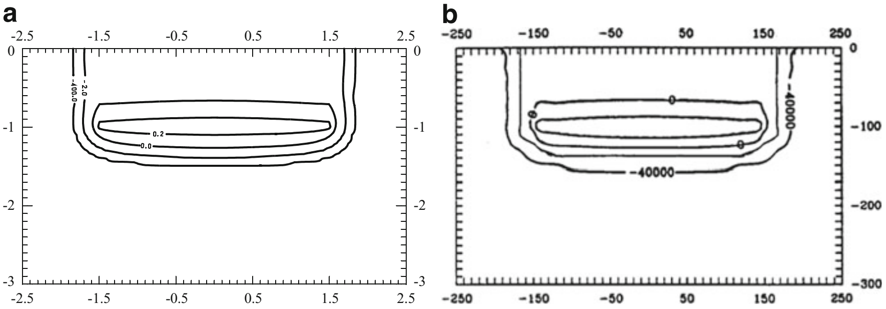


Fig. 10.30 Simulated pressure head contours ψ at $t = 1$ d: (**a**) present results, heads and lengths in (m) and (**b**) Kirkland et al.'s results [311], heads and lengths in (cm)

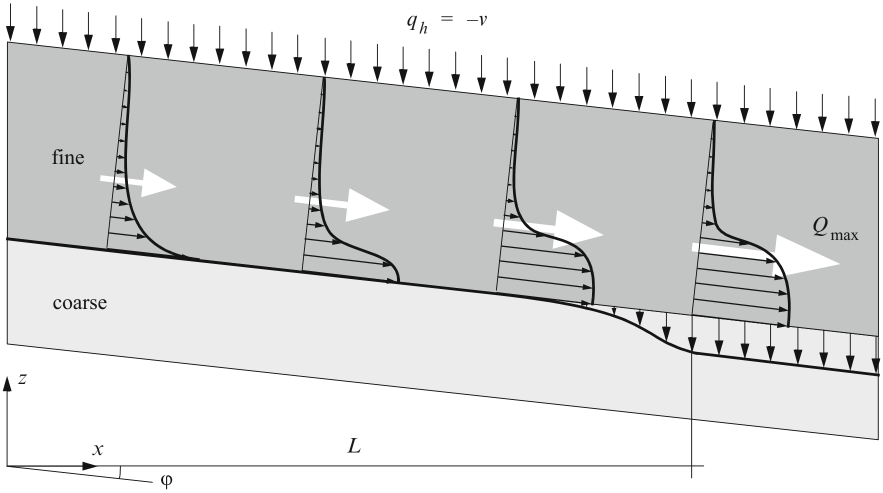


Fig. 10.31 Diverting flow profiles in a capillary barrier (Modified from [298])

10.13.5 Capillary Barrier Modeling

In unsaturated flow conditions a *capillary barrier* can appear at the contact of a layer of fine soil overlying a layer of coarse soil occurring both in natural situations and engineered systems [298, 397, 448, 473, 558]. If the layer interface is tilted, water infiltrating in the fine layer will be diverted and flow laterally down the contact (Fig. 10.31). 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). For such a fine-over-coarse soil layer structure, the capillary forces are usually high and prevent the water infiltration v entering into the underlying coarse layer over a certain distance downslope, forming the impervious capillary barrier. With increasing distance downslope the saturation in the fine layer further raises due to infiltration v and it occurs that the capillary forces can no longer prevent a water influx into the coarse layer. The capillary barrier starts to release some water. With the further increasing distance downslope, more and more water is released vertically into the underlying coarse layer reaching a distance L. From there the lateral (diverting) flow cannot be increased anymore and all further infiltrating water percolates entirely into the coarse layer. This point at distance L defines the breakthrough of infiltrating water. The length L is denoted as *diversion length*, which qualifies the effectiveness of a capillary barrier. For distances downslope larger than L, a capillary barrier becomes ineffective. To achieve large diversion lengths L, sharply contrasting properties for the fine-over-coarse soil system are required.

The discharge $Q_{\max}$ represents the maximum amount of water that the capillary barrier can divert in the fine layer. It is determined from the vertical integration of the horizontal flux of water diverted until breakthrough [448]

$$Q_{\max} = \int_{z_1}^{z_2} q_x dz = K \tan\varphi \int_{\psi_2}^{\psi_1} k_r(\psi) d\psi \tag{10.138}$$

where ψ_2 is the pressure head at the top surface of the fine layer z_2, ψ_1 is the pressure head at the bottom of the fine layer z_1 (which is equal to the pressure head at the top of the coarse layer), φ is the dip of the layers (Fig. 10.31), K and k_r are the saturated hydraulic conductivity and the relative permeability, respectively, of the fine layer. For a constant infiltration v the diversion length L is simply the maximum discharge $Q_{\max}$ divided by the infiltration rate v, $L = Q_{\max}/v$, so that

$$L = \frac{K}{v} \tan\varphi \int_{\psi_2}^{\psi_1} k_r(\psi) d\psi \tag{10.139}$$

Ross [448] has derived a closed-form expression[8] for the diversion length L, which results in using an exponential relationship for k_r. For other parametric models (10.139) must be integrated numerically with appropriate BC's [558]. In general, the numerical simulation of capillary barriers presents a significant challenge.

[8] Using the exponential relationship (D.39) in the form of $k_r = e^{\alpha\psi}$ we can integrate (10.139) analytically. The BC's at the top and bottom of the fine layer are the following: At the top, the relative permeability is simply the infiltration rate v divided by the saturated hydraulic conductivity K of the fine layer, i.e., $k_r = v/K = \exp(\alpha\psi_2)$ so that $\psi_2 = \frac{1}{\alpha}\ln(\frac{v}{K})$. At the bottom of the fine layer we find the pressure head equal to the value at the top of the coarse layer in a similar relation: $k_r^\star = v/K^\star = \exp(\alpha^\star\psi_1)$ so that $\psi_1 = \frac{1}{\alpha^\star}\ln(\frac{v}{K^\star})$, where $K^\star$ and $\alpha^\star$ are the saturated hydraulic conductivity and sorptive number, respectively, of the coarse layer. Applying these BC's for ψ_2 and ψ_1 we find the analytical solution of (10.139) for the diversion length as [448]

$$L = K\frac{\tan\varphi}{v\alpha}\Big[\Big(\frac{v}{K^\star}\Big)^{\alpha/\alpha^\star} - \Big(\frac{v}{K}\Big)\Big].$$

Table 10.23 Parameters and conditions used for Webb's capillary barrier problem

Quantity	Symbol	Value	Unit
Domain measure (length; thickness)		100; 1	m
Dip	φ	5	%
Specific storage coefficient	S_o	10^{-4}	m^{-1}
Material properties of the two layers are listed in Table 10.24			
IC and BC's			
Initial condition (IC)	h_0	$-z$	m
Neumann-type BC at surface	q_h	-0.0048	$m\,d^{-1}$
Dirichlet-type BC along both bottom and right vertical boundaries	h_D	0	m
FEM			
Nonuniform 2D mesh of 1,472 quadrilaterals in variable thicknesses			
Initial time step size	Δt_0	10^{-3}	d
RMS error tolerance (FE/BE)	ϵ	10^{-3}	1
Simulation time period	t_{end}	100	d

Table 10.24 Material properties for Webb's capillary barrier problem [558] by using the van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26)

Layer	Thickness (m)	ε (1)	s_r (1)	α (m^{-1})	n (1)	K ($m\,s^{-1}$)
Fine (upper)	0.5	0.39	0.3945	3.9	5.74	$2.1 \cdot 10^{-4}$
Coarse (lower)	0.5	0.42	0.0286	490	2.19	0.1

[a] Note that it is generally used: $s_s = 1, m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

The numerical schemes have to tackle large parameter contrasts, highly exaggerated and distorted geometries as well as dry IC's.

10.13.5.1 Webb's Problem

Oldenburg and Pruess [397] 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 [448] was achieved. Webb [558] could improve the steady-state results by using an upstream weighting technique agreeing well with Ross' analytical prediction [448].

We use Webb's capillary barrier problem [558] to study the capability of the present approaches. 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 and conditions used for simulating Webb's two-layer problem are summarized in Tables 10.23 and 10.24. The infiltration rate v at the surface of the domain is $0.0048\,m\,d^{-1}$. The left boundary is impervious and the right and bottom boundaries allow for drainage. This can be done in several ways. In consideration of the extreme parameter situation of the fine and coarse layers (cf. Table 10.24) we found a reasonable convergence

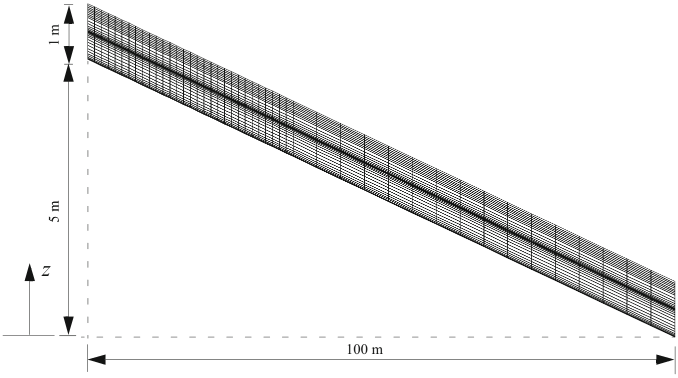


Fig. 10.32 Model domain and mesh (1,472 quadrilaterals with 1,551 nodes) for Webb's capillary barrier problem [558] (vertical exaggeration 10:1)

behavior for a Dirichlet-type BC, where the hydraulic head h is imposed. Since the $\alpha-$parameter of the coarse layer is very large the influence of the location of the water table (the $\psi = 0$ condition) cannot be significant. It is thus sufficient to set the water table at the right lower corner of the domain (at $z = 0$) and prescribe a $h = 0$ Dirichlet BC along the bottom and the right boundaries. In accordance with this BC a corresponding IC is assumed possessing a vertical linear distribution of $h_0 = -z$ in the range from 0 to -6 m. This results in averaged initial saturations s_0 which are very close to the residual saturations s_r (cf. Table 10.24). The model domain is appropriately discretized in quadrilateral elements as displayed in Fig. 10.32. At the layer contact the element thickness is 0.005 m and gradually increases with the distance from the interface.

Figure 10.33 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. 10.33. The diversion is maintained up to a certain distance, the diversion length L, past which an amount of water equal to the infiltration rate v enters the coarse layer.

A comparison of the above results with Ross' analytical formula [448] and the numerical results obtained by Webb [558] can be expressed as a function of the leakage/infiltration ratio. The theoretical value of the diversion length L determined from Ross' formula (10.139) is 32.6 m for the VGM present parameters (note, Webb [558] computed 33.2 m). As evidenced in Fig. 10.34 there is a good qualitative and quantitative agreement between the analytical and the numerical results. Note 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. 10.34 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.

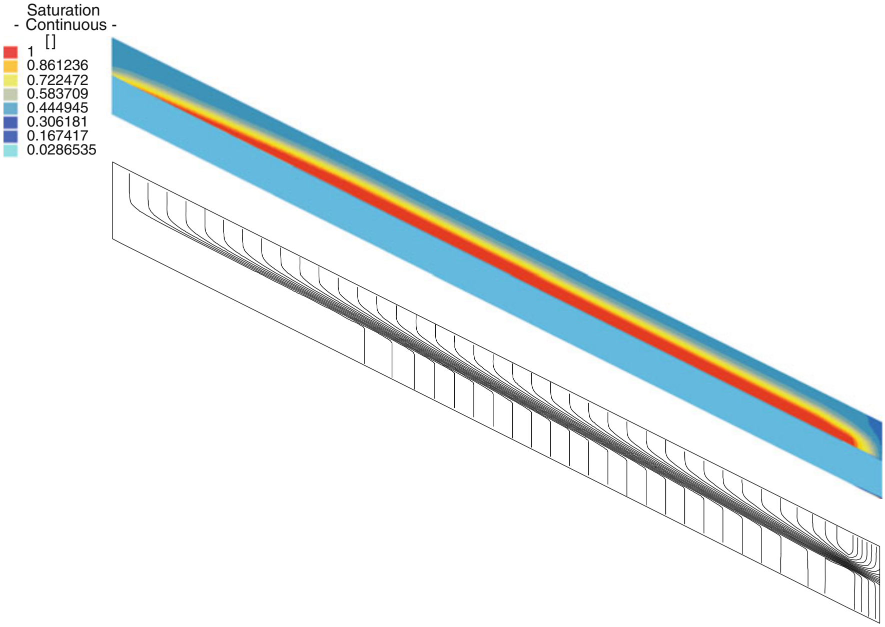


Fig. 10.33 Computed saturation and streamline pattern at $t = 100\,\text{d}$ for Webb's capillary barrier [558] (vertical exaggeration 10:1)

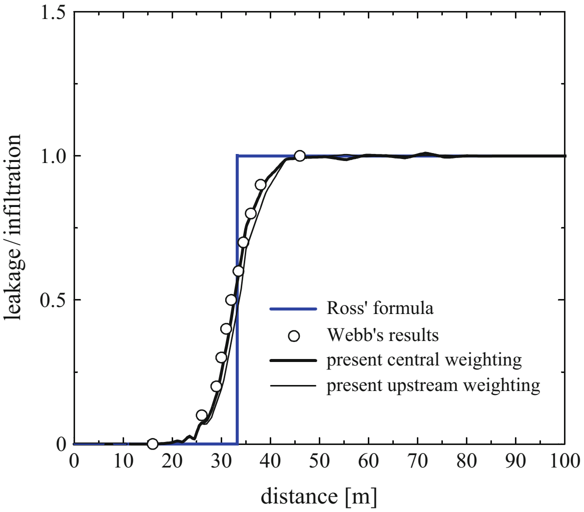


Fig. 10.34 Leakage/infiltration ratio in the coarse layer for both central and upstream weighting compared to Ross' analytical formula (10.139) and Webb's numerical results [558]. Diversion length of $L = 32.6\,\text{m}$ results from Ross' formula

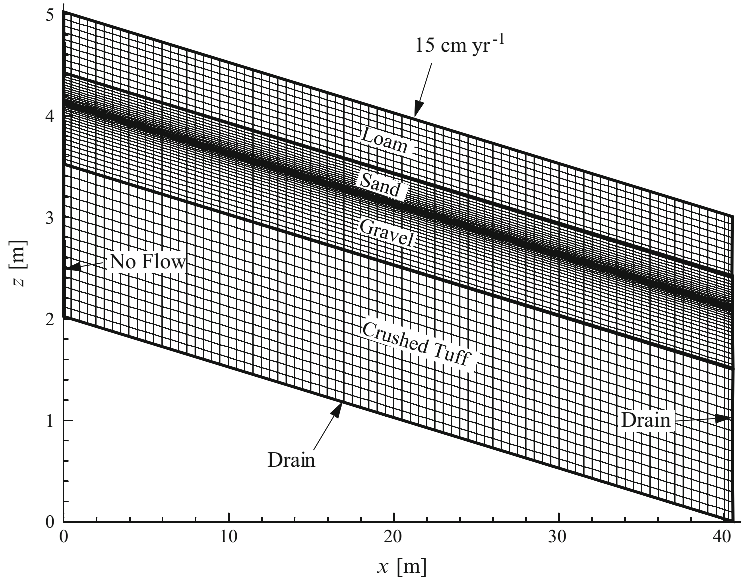


Fig. 10.35 Capillary barrier model domain with used mesh consisting of 82 × 61 isoparametric bilinear elements (vertical exaggeration 5:1)

It is to be mentioned that the specific advantages of the PVST 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 $h-$based form becomes an effective alternative. We obtained the above solutions for the $h-$based form of the Richards' equation, using the FE/BE predictor-corrector time stepping scheme with the Picard iteration method.

10.13.5.2 Forsyth and Kropinski's Problem

A numerically challenging capillary barrier problem was considered by Forsyth and Kropinski [166]. The problem is described in Fig. 10.35 and Table 10.25. The material properties and the initial pressure conditions for the different layers are given in Table 10.26. As indicated the IC's enforce very dry soils. The infiltration rate at the surface of the cross-sectional domain is 15 cm year^{-1}. The mesh is shown in Fig. 10.35 consisting of 82 × 61 quadrilateral linear elements with 5,146 nodes. As seen the element size is highly variable in the vertical direction. At the sand-gravel interface the elements have a thickness as small as 0.002 m. The left vertical boundary is considered impervious. To model free drainage at the bottom and the right vertical boundary of the domain, proper gradient-type BC's of $q_h^{\nabla} = K$ are imposed, where K are the hydraulic conductivities of the layers (Table 10.26) at the corresponding boundary sections.

Table 10.25 Parameters and conditions used for Forsyth and Kropinski's capillary barrier problem

Quantity	Symbol	Value	Unit
Domain in Fig. 10.35 (length; thickness)		40.5; 3	m
Dip	φ	5	%
Specific storage coefficient	S_o	0	m^{-1}
Material properties of the four layers are listed in Table 10.26			
IC and BC's			
Initial condition (IC) of pressure p	p_0	Table 10.26	kPa
Neumann-type BC at surface	q_h	$-4.11 \cdot 10^{-4}$	$\mathrm{m\,d}^{-1}$
Gradient-type BC along both bottom and right vertical boundaries	q_h^{∇}	K (Table 10.26)	$\mathrm{m\,d}^{-1}$
FEM			
Nonuniform 2D mesh of 82×61 quadrilaterals, highly variable in vertical direction			
Initial time step size	Δt_0	10^{-5}	d
Maximum error tolerance (FE/BE)	ϵ	10^{-3}	1
Residual error tolerance	$\epsilon_2^{\star}$	10^{-4}	$\mathrm{m}^3\,\mathrm{d}^{-1}$
Simulation time period	t_{end}	30	years

Table 10.26 Material properties and IC's for Forsyth and Kropinski's capillary barrier problem [166] by using the van Genuchten-Mualem (VGM) parametric model[a] (D.3), (D.26)

Layer	Thickness (m)	ε (1)	s_r (1)	α (m^{-1})	n (1)	K ($\mathrm{m\,s}^{-1}$)	p_0[b] (kPa)
Loam	0.6	0.452	0.0752	4.3	1.246	$1.668 \cdot 10^{-5}$	-10^6
Sand	0.3	0.345	0.046	6.34	1.53	$6.573 \cdot 10^{-5}$	-10^6
Gravel	0.6	0.419	0.074	469	2.57	$3.502 \cdot 10^{-3}$	-30
Crushed tuff	1.5	0.345	0.032	1.43	1.506	$2.776 \cdot 10^{-6}$	$-6 \cdot 10^{10}$

[a] Note that it is generally used: $s_s = 1, m = 1 - \frac{1}{n}$ and $\sigma = \frac{1}{2}$

[b] Pressure p is related to pressure head ψ by (3.259): $p = \psi \rho_0 g$, where $\rho_0 = 10^3\,\mathrm{kg\,m}^{-3}$ and $g = 9.81\,\mathrm{m\,s}^{-2}$

Due to the extremely dry IC's the PVST is the favorable solution strategy. Central and upstream weighting are applied. However, the FE/BE predictor-corrector time stepping requires a large number of time steps to simulate the complete 30-year time period (up to 130,000 time steps for central weighting by using the maximum error tolerance of Table 10.25) caused by the demand for bounding the temporal discretization error. This is overdone if the major interest is only in the final solution at steady-state. For this need a target-based scheme appears to be more useful for which only about 5,000 time steps (with about 10^4 total Newton steps) are necessary [141].

The present simulation results confirm Forsyth and Kropinski's findings [166]. The computed saturation distributions are displayed for three specific contour levels in Fig. 10.36 for the central weighting and in Fig. 10.37 for the upstream weighting. Some details are depart from Forsyth and Kropinski's simulations. It can be assumed that most of them is caused by different BC's. 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,

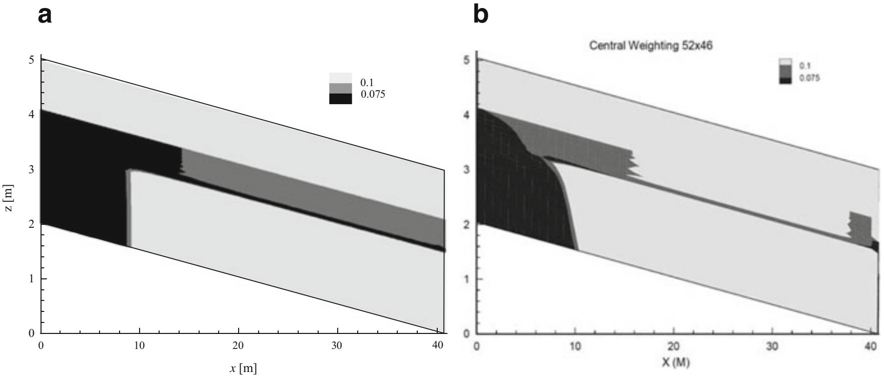


Fig. 10.36 Saturation patterns simulated with central weighting at $t = 30$ years: (**a**) present solution for the 82×61 mesh and (**b**) Forsyth and Kropinski's results [166] for a 52×46 mesh

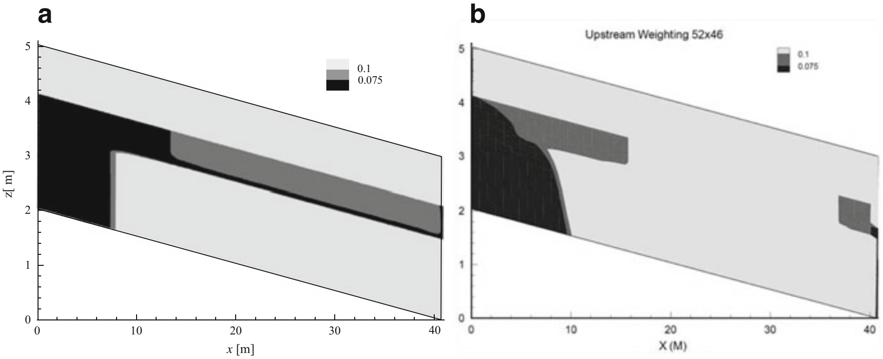


Fig. 10.37 Saturation patterns simulated with upstream weighting at $t = 30$ years: (**a**) present solution for the 82×61 mesh and (**b**) Forsyth and Kropinski's results [166] for a 52×46 mesh

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. 10.35. For the central weighting (Fig. 10.36) we note a jagged saturation profile which disappears for upstream weighting (Fig. 10.37). 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. 10.36b).

The computed streamline pattern in Fig. 10.38 illustrates the effect of the capillary barrier at the sand-gravel material interface. The streamlines reveal that the diversion length L is larger than Forsyth and Kropinski's estimation with 10 m. Actually, the velocity distribution along the bottom of the tuff layer indicates a leakage increase from zero at about 10 m to the infiltration rate at about 25 m, as depicted in Fig. 10.39. This 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

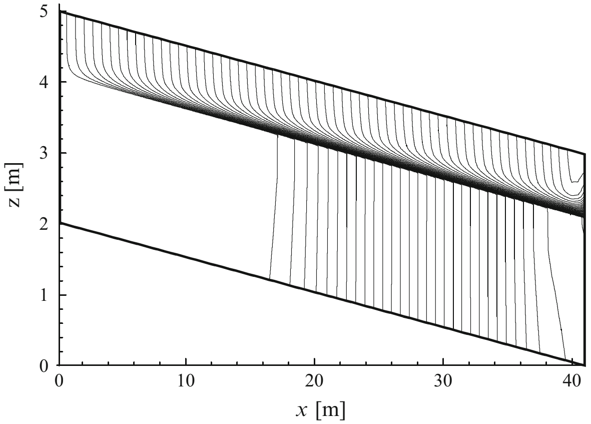


Fig. 10.38 Streamline pattern simulated with central weighting at $t = 30$ years

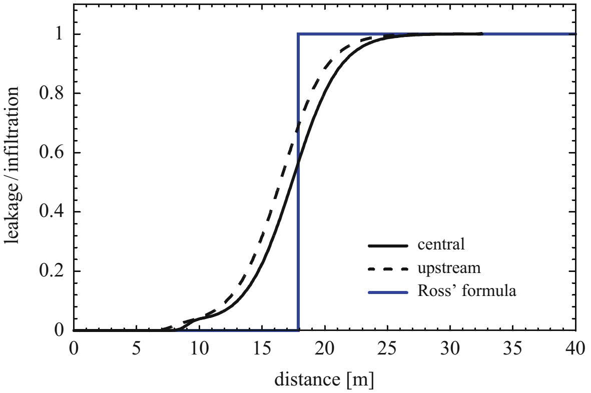


Fig. 10.39 Leakage/infiltration ratio in the tuff layer. Analytical diversion length results in $L = 17.9\,\mathrm{m}$

(10.139) using the above van Genuchten parameters for the sand and gravel zones (Table 10.26) gives a diversion length of $L = 17.9\,\mathrm{m}$. This value is in good agreement with the present numerical simulations as seen in Fig. 10.39.

10.13.6 Dam Seepage Problem

Seepage through and under embankment dams is one of the standard tasks in finite element modeling, in which unsaturated-saturated flow regimes have to be analyzed, e.g., [329, 383]. If fluctuations in reservoir level are present the transient seepage through the dam is highly dependent on conditions of the unsaturated zone. To study the hydrodynamics of dam drainage and remedial sealing measures let us consider

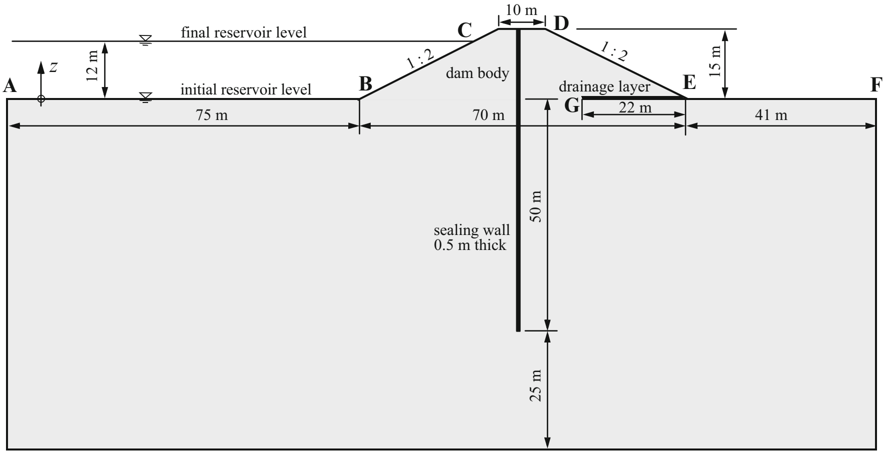


Fig. 10.40 Reservoir with embankment dam and selected underground

the example as displayed in Fig. 10.40. The dam and its underground consist of a homogeneous isotropic sandy material. We simulate the seepage process when the water reservoir becomes flooded. The reservoir is initially empty. The water level in the reservoir is now raised up to 12 m during 5 days, afterwards the level in the reservoir remains at the constant elevation of 12 m, without accretion. The dam is equipped with a horizontal drain at the dam toe. It is assumed that this drain filter is insufficiently operational. This is mimicked by a Cauchy-type BC which admits only a limited drainage capacity controlled via the transfer rate $\Phi_h = \Phi_h^{\text{out}} \geq 0$. At the complete downstream slope of the embankment and the ground surface a seepage face BC is imposed which allows a free drainage of water. To reduce the seepage through and under the dam a partial sealing wall is considered as shown in Fig. 10.40. We compare the transient seepage process without and with the partial sealing wall. The used model parameter and conditions are summarized in Tables 10.27 and 10.28. Unspecified BC's represent no-flow conditions. Estimated material parameters are used for the modified van Genuchten parametric model, where the relative permeability $k_r = s_e^\delta$ can be typically accepted as a linear relationship by using $\delta = 1$ for the present class of seepage problems in which the major interest is in determining the location of the free surface.

The simulation of the transient seepage problem can suitably based on the standard $h-$form of the Richards' equation, where the adaptive AB/TR predictor-corrector time stepping scheme is preferred. The used finite element mesh shown in Fig. 10.41 is appropriately refined in the dam body, at the sealing wall and along the seepage face. The mesh is designed to handle different lengths of the sealing wall. Figure 10.42 exhibits the simulated free-surface development, i.e., the advance of the zero-pressure head ($\psi = 0$) contour line, for the case without and with the partial sealing wall. In both cases the flow reaches steady-state conditions after about 50 days. The total discharge through and under the dam is predicted

Table 10.27 Parameters and conditions used for the dam seepage problem

Quantity	Symbol	Value	Unit
Domain shown in Fig. 10.40			m
Specific storage coefficient	S_o	10^{-4}	m^{-1}
Material properties are listed in Table 10.28			
IC and BC's			
Initial condition (IC)	h_0	0	m
Dirichlet-type BC at $\widehat{ABC}$[a]	$h_D(t)$	$\begin{cases} 2.4 \cdot t & (0 \leq t \leq 5 \text{ d}) \\ 12 & (t > 5 \text{ d}) \end{cases}$	m
Seepage face BC at $\widehat{DEF}$[a]	$\begin{cases} h = z \\ Q_{n_h} < 0 \end{cases}$		m
Cauchy-type BC at $\widehat{GE}$[a]	h_D	0.5	m
	Φ_h^{out}	$5 \cdot 10^{-2}$	d^{-1}
	Φ_h^{in}	0	d^{-1}
FEM			
Unstructured 2D mesh of 20,451 triangles, refined at dam body and sealing wall			
Initial time step size	Δt_0	10^{-3}	d
Maximum error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	1	year

[a] Defined in Fig. 10.40

Table 10.28 Material properties for the dam seepage problem by using the modified van Genuchten parametric model[a] (D.3), (D.33)

Zone	ε (1)	s_r (1)	α (m^{-1})	n (1)	m (1)	δ (1)	K ($\mathrm{m\,s}^{-1}$)
Dam body/underground	0.45	0.3	1.3	2.2	0.545	1	$2 \cdot 10^{-5}$
Sealing wall	0.55	0.4	0.8	1.8	0.444	1	10^{-8}

[a] Note that it is generally used: $s_s = 1$

by 18.91 $\mathrm{m}^3\,\mathrm{d}^{-1}/\mathrm{m}$ at steady-state for the case without the sealing wall. With the partial sealing wall it reduces to 7.79 $\mathrm{m}^3\,\mathrm{d}^{-1}/\mathrm{m}$. Figure 10.43 illustrates the general steady-state flow pattern in the presence of the partial sealing wall in form of the computed flow net.[9] It clearly indicates the effect of the sealing wall achieving a reduced seepage loss and flow gradients at the downstream side of the dam.

[9] In 2D and under steady-state conditions equipotential lines are given by the interval of hydraulic head Δh. The interval of streamlines (actually, interval of the streamfunction, cf. Sect. 2.1.11) $\Delta\Psi$ is determined from

$$\Delta\Psi = K \frac{\Delta h}{\Delta l} \Delta q$$

where Δl is the distance between two neighboring equipotential lines and Δq is the width of the stream tube. A flow net can be constructed if setting $\Delta l = \Delta q$ so that streamlines and equipotential lines form 'curvilinear squares'. For such a flow net configuration it is

$$\Delta\Psi = K \Delta h.$$

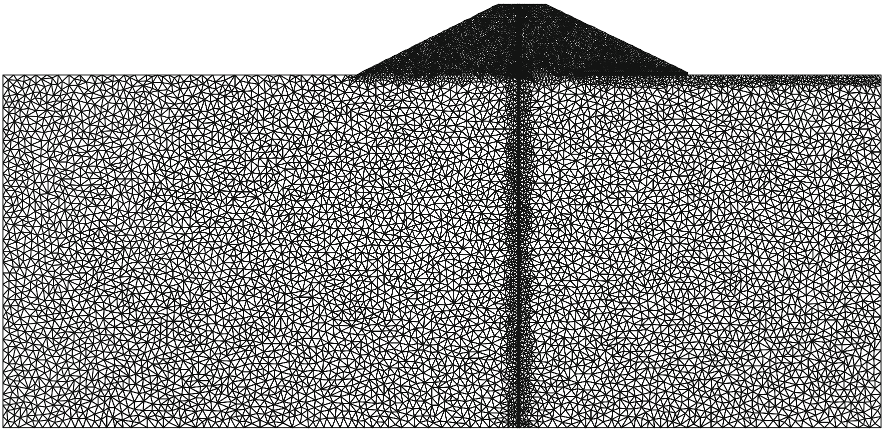


Fig. 10.41 Unstructured triangular mesh used for modeling the dam seepage problem in the 2D cross section ($N_E = 20,451$, $N_P = 10,418$)

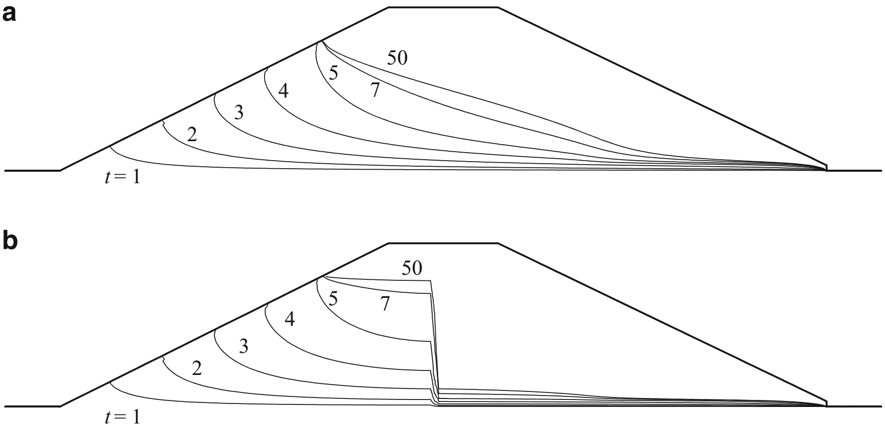


Fig. 10.42 Advance of free surface ($\psi = 0$) in accordance with the raised water level in the reservoir: (**a**) without and (**b**) with sealing wall, times t in (d)

10.13.7 On Draining and Flooding

In many applications the modeling of drainage or flooding processes in regional 3D phreatic aquifer systems are of specific concern. Typical examples refer to the impact of drainage from mine operations and flooding of abandoned open-pit mines. It is interesting to know how true unsaturated-saturated modeling approaches are appropriate to simulate such type of large-scale flow problems in comparison to the classic free-surface groundwater modeling strategies of Sect. 9.5. For this purpose let us study the three following examples which are sufficiently generic in this field.

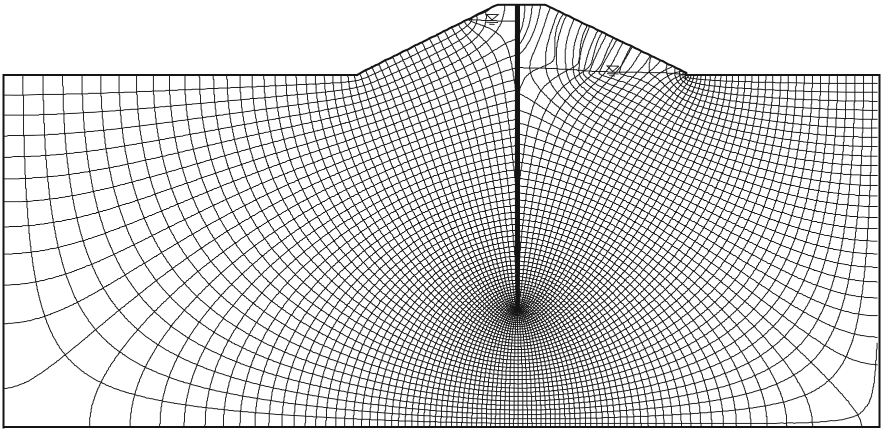


Fig. 10.43 Computed flow net at steady-state for the dam seepage with partial sealing wall (used intervals $\Delta h = 0.1\,\text{m}$ and $\Delta\Psi = 0.1728\,\text{m}^2\,\text{d}^{-1}$)

10.13.7.1 Vachaud et al.'s Problem: Drainage Experiment

Vachaud et al. [529] 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 [216] and elsewhere [382, 474]. A half drain-spacing with a length of 3 m and a height of 2 m is considered (Fig. 10.44). Initially, the water level in the box is at $z = h_0$ and the system is under hydrostatic equilibrium with $\psi_0 = h_0 - z$. The soil is assumed to be isotropic with a saturated conductivity of $1.11 \cdot 10^{-4}\,\text{m}\,\text{s}^{-1}$. The Haverkamp parametric model (D.11), (D.36) is used for the unsaturated soil. The initial hydraulic 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 has to be determined in the solution process. In Vachaud et al.'s experiment the drainage process has been performed without any infiltration ($v = 0$ on top, Fig. 10.44). Accordingly, the water table descends continuously up to reaching the water level h_w of the ditch. In contrast, Gureghian's results [216] are based on an infiltration rate given with $v = 0.384\,\text{m}\,\text{d}^{-1}$ so that his solution approaches to a non-horizontal water table in time. The parameters and conditions used in the present simulations are summarized in Table 10.29.

We use the standard form of the Richards' equation with the FE/BE predictor-corrector method. The 2D domain is discretized by only 640 quadrilateral elements as shown in Fig. 10.45 with the major BC's. Figure 10.46 compares the present numerical results with Vachaud et al.'s experimental data. As seen the agreement is quite well. A comparison of the hydraulic head contours, the water table and capillary fringe at time of hour is presented in Fig. 10.47 between the present solutions and Gureghian's results. The agreement is reasonable. Differences appear for the upper head contours which obviously result from the different description of the infiltration BC.

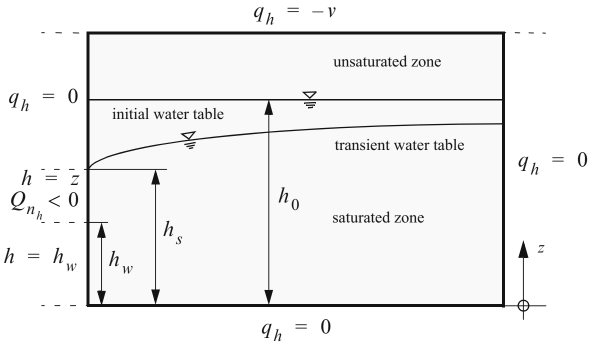


Fig. 10.44 Sketch of Vachaud et al.'s drainage experiment [529]: geometry and BC's

Table 10.29 Parameters and conditions used for Vachaud et al.'s problem (Fig. 10.44)

Quantity	Symbol	Value	Unit
Box measure (length; height)		3; 2	m
Saturated conductivity	K	$1.11 \cdot 10^{-4}$	$\mathrm{m\,s^{-1}}$
Porosity	ε	0.3	1
Specific storage coefficient	S_o	10^{-4}	$\mathrm{m^{-1}}$
Maximum saturation	s_s	1	1
Residual saturation	s_r	0	1
Haverkamp parametric model (D.11), (D.36)			
Fitting coefficient	α	0.063396	m
Fitting coefficient	A	$3.6 \cdot 10^{-4}$	m
Fitting exponent	β	2.9	1
Fitting exponent	B	4.5	1
IC and BC's			
Initial condition (IC)	h_0	1.45	m
Neumann-type BC at top (infiltration)	q_h	−0.384	$\mathrm{m\,d^{-1}}$
Seepage face BC at ditch (within $0.75\,\mathrm{m} < z \leq 1.45\,\mathrm{m}$)	$\begin{cases} h = z \\ Q_{n_h} < 0 \end{cases}$		m $\mathrm{m^3\,d^{-1}}$
Dirichlet-type BC at ditch (within $0\,\mathrm{m} \leq z \leq 0.75\,\mathrm{m}$)	h_w	0.75	m
FEM			
2D mesh of quadrilateral elements	displayed in Fig. 10.45		
Mesh spacing	$\Delta x\ /\ \Delta z$	15 / 5.83 − 6.875	cm
Number of quadrilateral elements	N_{E}	$20 \times 32 = 640$	
Number of mesh nodes	N_{P}	$21 \times 33 = 693$	
Initial time step size	Δt_0	10^{-3}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Simulation time period	t_{end}	5	h

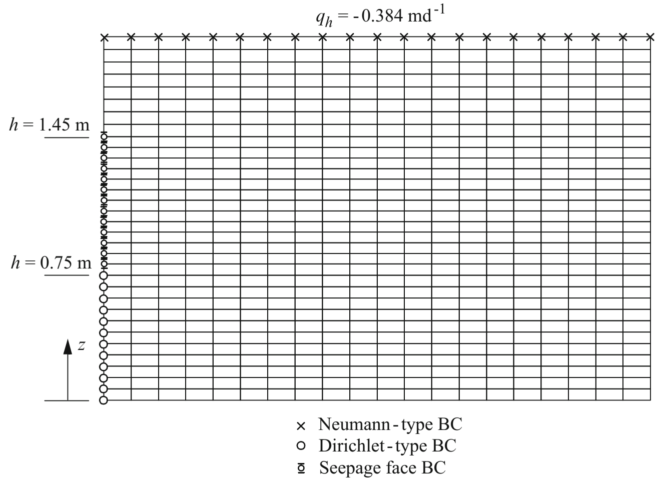


Fig. 10.45 Used quadrilateral finite element mesh with BC's

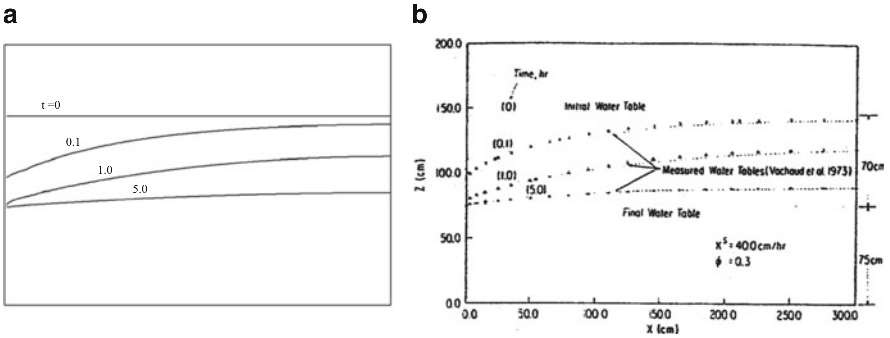


Fig. 10.46 Descending water table of the drainage experiment: (**a**) simulated free-surface locations and (**b**) Vachaud et al.'s measured water tables given in [216], times in hours

10.13.7.2 Free Drainage of a Thick Porous Block

The prototypical 3D example is shown in Fig. 10.48. The data and conditions used in the simulations are summarized in Table 10.30. The vertical drainage of a thick aquifer is modeled by a porous block with a base of 1 square meter ($A = 1\,\mathrm{m}^2$) and height of 100 m. The block is initially saturated, the water table h is on top at $h_0 = 100\,\mathrm{m}$. Now, the domain begins to be freely drained and the water table draws down continuously in time t due to the dewatering at the bottom of the block enforced by a significantly deeper water level with $h_D = 0\,\mathrm{m}$. The flow is purely gravitationally driven and basically 1D in the vertical direction. Accordingly, the theoretical drainage capacity (discharge) Q at the bottom must be constant at larger times, viz.,

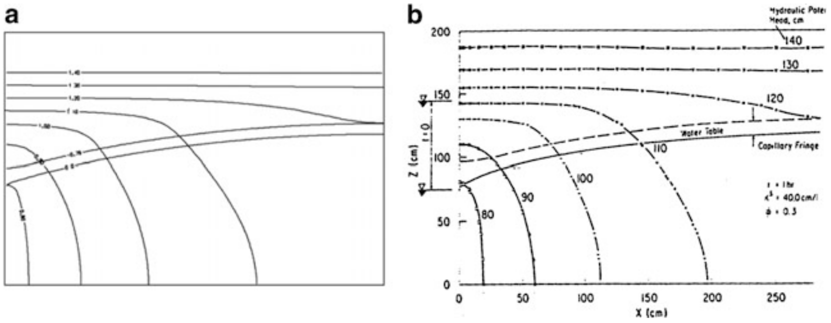


Fig. 10.47 Hydraulic head contour and water table location at $t = 1$ h: (**a**) present results and (**b**) Gureghian's solutions [216]

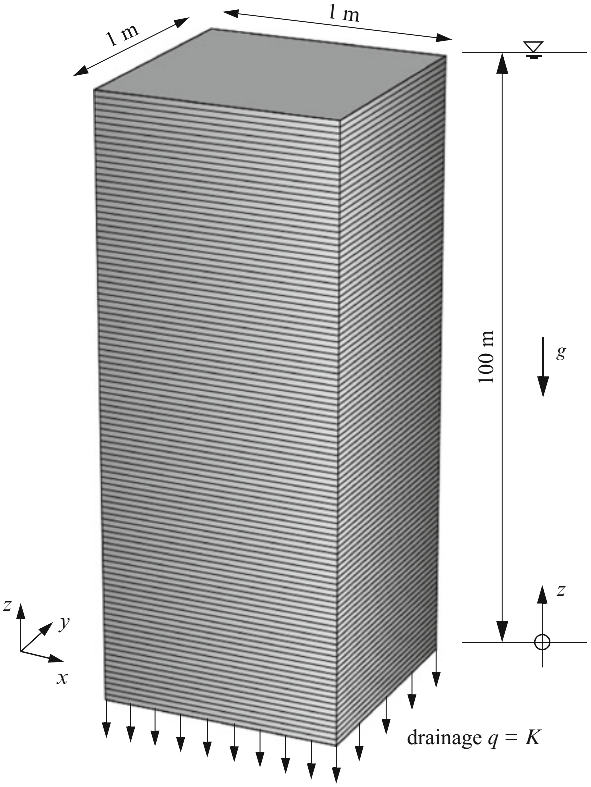


Fig. 10.48 Study domain of the thick porous block and used 3D mesh consisting of 100 hexahedral elements with $\Delta z = 1$ m (vertical exaggeration 0.03:1)

Table 10.30 Parameters and conditions used for the thick porous block drainage

Quantity	Symbol	Value	Unit
Block height	L	100	m
Block base	A	1	m^2
Block volume	V	100	m^3
Total drainable water content	V^w	20	m^3
Saturated conductivity	K	10^{-5}	$\mathrm{m\,s^{-1}}$
Porosity	ε	0.25	1
Specific storage coefficient	S_o	10^{-8}	m^{-1}
Maximum saturation	s_s	1	1
Residual saturation	s_r	0.2	1
Specific yield	ε_e	0.2	1
Modified van Genuchten parametric model (D.3), (D.33)			
Pore size distribution index	n	2	1
Fitting coefficient	α	10	m^{-1}
Fitting exponent	m	0.5	1
Fitting exponent	δ	1	1
IC and BC			
Initial condition (IC)	h_0	100	m
Dirichlet-type BC at bottom	h_D	0	m
FEM			
Vertical space increment	Δz	1	m
Initial time step size	Δt_0	10^{-8}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Simulation time period	$t_{\text{end}} = t_{\max}$	23.15	d

$$Q = K\,A \qquad (0 < t \le t_{\max}) \tag{10.140}$$

where $t_{\max}$ is the maximum time reached when the domain is fully drained out. The initially (total) drainable water content of the block is $V^w = \varepsilon_e\,V$, where V is the block volume and $\varepsilon_e = \varepsilon(1 - s_r)$ is the specific yield (3.296). Thus, the duration of the drainage results in

$$t_{\max} = \frac{V^w}{Q} = \frac{\varepsilon_e V}{KA} = \frac{\varepsilon_e h_0}{K} \tag{10.141}$$

which gives $t_{\max} = 23.15\,\mathrm{d}$ by using the parameters listed in Table 10.30. The vertical drainage process is governed by the 1D flow equation for the (free-surface) water table h as

$$\varepsilon_e \frac{\partial h}{\partial t} = -\frac{\partial}{\partial z}(h\,q) \tag{10.142}$$

where $q = Q/A = K$ is the drainage rate. Assuming $q = \text{const}$ we obtain the analytical solution for the water table $h(z,t)$ from (10.142) as

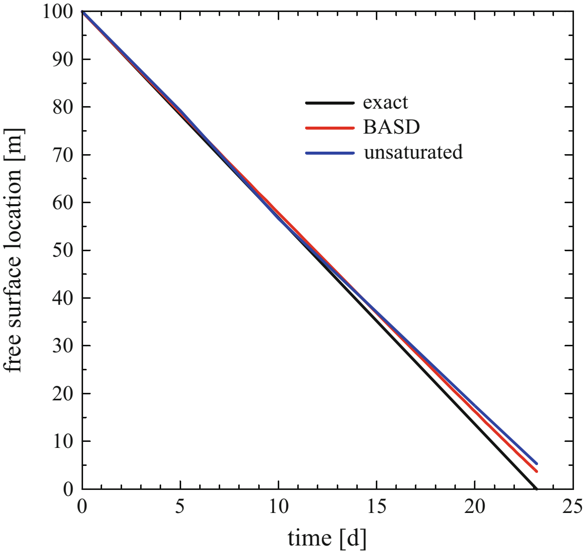


Fig. 10.49 Descent of free surface in time ($0 \le t \le t_{\max}$). Classic free-surface groundwater model results obtained with moving mesh BASD technique and unsaturated-saturated flow computation based on the modified VG parametric model in comparison to the exact solution (10.143). In the unsaturated-saturated flow model the free surface location is defined for a small negative pressure head of $\psi = -0.15\,\text{m}$

$$h(z,t) = \frac{1}{\varepsilon_e}\big(\varepsilon_e z - Kt\big) \quad \text{with} \quad z = h_0 \tag{10.143}$$

which indicates that the water table h must fall simply linearly from h_0 to zero within the time range $0 < t \le t_{\max}$.

The simulations are performed with the uniform 3D mesh shown in Fig. 10.48 by using both the standard $h-$formulation of unsaturated-saturated flow model and the classic free-surface groundwater strategy based on moving mesh BASD technique.[10] Adaptive FE/BE predictor-corrector time stepping is preferred in both approaches. The classic free-surface computation gives excellent results in comparison to the exact solution (10.143) as shown in Fig. 10.49 for the linear decrease of the water table and in Fig. 10.50a for the linear rise of the accumulated drainage volume $V^w(t) = A\int_t q(t)dt$ in time, where the drainage rate $q(t)$ is actually measured at the bottom face of the porous block during the simulation. We obtain exactly $V^w(t = t_{\max}) = 20\,\text{m}^3$ when the block is drained out. It is important to note that the linearly falling water table according to (10.143) can only occur at larger times after the gradient of the pressure head $\partial\psi/\partial z$ disappears in the saturated zone of the block. Due to the initial hydrostatic head condition of $h_0 = 100\,\text{m}$ and the abrupt

[10] The present example is easily solvable for classic free-surface flow modeling with moving mesh (cf. Sect. 9.5.3), even with only a small number of elements. Contrarily, the classic free-surface modeling strategy with fixed mesh and pseudo-unsaturated conditions (cf. Sect. 9.5.4) will not give reasonable results for such type of a vertically dominant drainage because the free-surface BC assigned unmovably to the upper element slice becomes ineffective when all underlying elements fall dry in time.

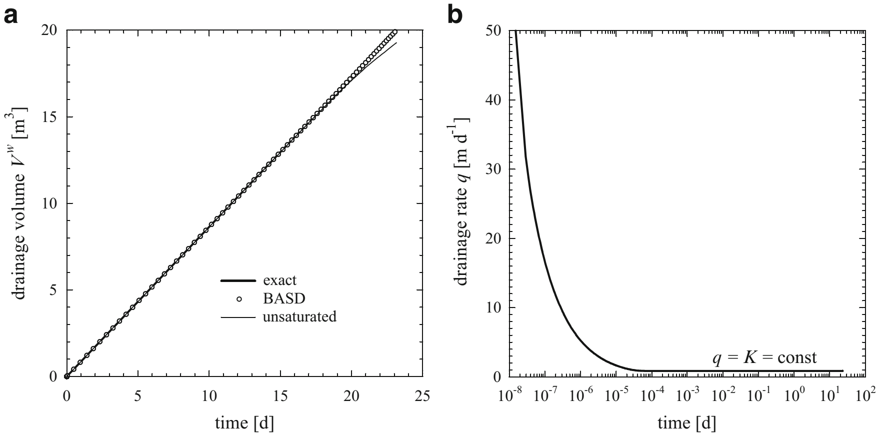


Fig. 10.50 (**a**) Accumulated drainage volume $V^w(t) = A\int_t q(t)dt$ in time $(0 \leq t \leq t_{\max})$. Classic free-surface groundwater model results obtained with moving mesh BASD technique and unsaturated-saturated flow computation based on the modified VG parametric model in comparison to the exact solution $V^w(t) = AKt$. (**b**) The behavior of $q(t)$ magnified at small times t

head gradient enforced by the drainage BC $h_D = 0\,\text{m}$ at the bottom face of the porous block, a little moment must elapse to equilibrate $\partial\psi/\partial z \approx 0$ throughout the (initially saturated) block, afterwards the constant free drainage establishes at the bottom and the descent of free surface on top starts. The duration of the initial pressure redistribution is controlled by the ratio K/S_o (with $S_o > 0$) and takes only about 2 s for the chosen parameters. During this short time period the drainage rate q approaches from a large (theoretically infinite) value to the constant $q = K$ as shown in Fig. 10.50b. Note that one could skip this phase if the simulation starts immediately with $\psi_0 = 0$ (or $h_0 = z$) from beginning.

The simulation results are in good agreement with the exact predictions for both modeling strategies as shown in Figs. 10.49 and 10.50. Slight deviations are revealed towards the end of drainage at $t_{\max}$. The moving mesh strategy must avoid a mesh collapsing when the free surface approaches the bottom, while the unsaturated-saturated model is influenced by the capillary rise induced by the lower BC when the block becomes fully unsaturated. While the location of the free surface is naturally given by the top boundary of the moving mesh, the free surface location cannot be easily taken from $h-$values at nodes in the unsaturated zone for the unsaturated-saturated model of a invariable mesh because the pressure head in the unsaturated zone is only determined by the retention curve and therefore strongly dependent on the used retention parameters. This becomes clear if seeing the obtained history of the pressure profiles for the unsaturated-saturated model in Fig. 10.51. Since the pressure head ψ is zero in the saturated zone, an equivalent free surface is to be defined at the transition to the unsaturated zone with a certain small negative pressure head. Actually, a small negative pressure head of $\psi = -0.15\,\text{m}$ is taken. Note that the sharper the retention curve is chosen, the better the transition

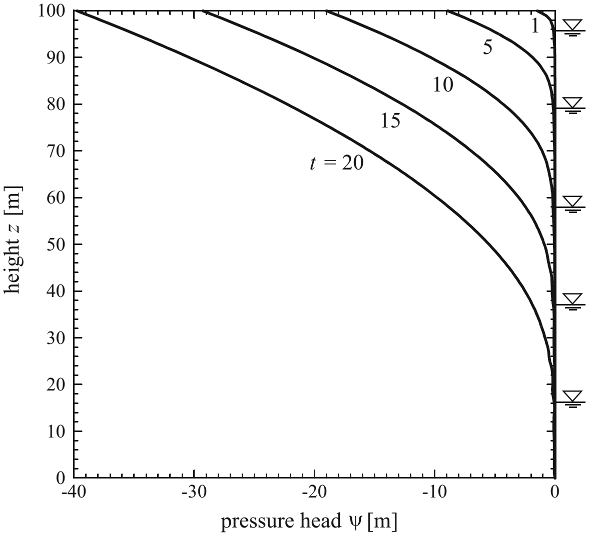


Fig. 10.51 Vertical pressure head profiles at selected times t computed with the unsaturated-saturated flow model. The free surface is defined at a small negative pressure head of $\psi = -0.15\,\text{m}$, times t in (d)

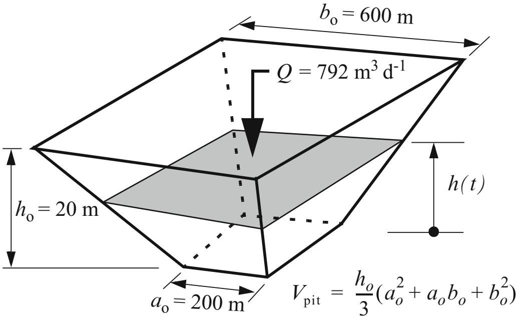


Fig. 10.52 Sketch of the pyramidal pit

from the saturated to the unsaturated zone can be identified. On the other hand, however, a sharp retention curve requires a higher spatial resolution and increases the computational effort.

10.13.7.3 Pit Flooding Test Case

Let us consider a simplistic open-pit mine with a geometry as shown in Fig. 10.52. The pyramidal pit body with a volume V_{pit} of $3,466,666.67\,\text{m}^3$ will be flooded by a constant water discharge Q of $792\,\text{m}^3\,\text{d}^{-1}$. The pit is initially dry and we assume for this in-pit domain a hydraulic conductivity of $K = 100\,\text{m}\,\text{s}^{-1}$, a specific yield of $\varepsilon_e = 1$ and a storage coefficient of $S_o = 0$. The surrounding porous body is considered very low permeable, actually, we assign $K = 10^{-9}\,\text{m}\,\text{s}^{-1}$, $\varepsilon_e = 0$ and $S_o = 10^{-4}\,\text{m}^{-1}$. As the solution the filling curve $h = h(t)$ is to be determined.

Table 10.31 Parameters and conditions used for the pit flooding problem

Quantity	Symbol	Value	Unit
Domain measure (width; depth; height)		800; 800; 21.5	m
Geometry of the mined-out pit is shown in Fig. 10.52			
Material properties of the in-pit domain and the surround are listed in Table 10.32			
IC and BC			
Initial condition (IC)[a]	h_0	10^{-4}	m
Multilayer well BC at pit center	Q_w	792	$m^3\,d^{-1}$
FEM			
3D mesh of 32 × 32 × 2 brick elements for the classic free-surface BASD model			
3D mesh of 32 × 32 × 21 brick elements for the unsaturated-saturated model			
Initial time step size	Δt_0	10^{-4}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	12	years

[a] Elevation $z = 0$ is defined at the pit base

Table 10.32 Material properties for the pit flooding problem. Unsaturated-saturated model uses the modified van Genuchten parametric model[a] (D.3), (D.33). Classic free-surface BASD model with moving mesh uses the specific yield ε_e^{b}

Subdomain	S_o (m^{-1})	K ($m\,s^{-1}$)	ε (1)	s_r (1)	α (m^{-1})	n (1)	m (1)	δ (1)
Pit	0	10^2	1	0	(10; 1)	2	0.5	1
Surround	10^{-4}	10^{-9}	0	0	1	2	0.5	1

[a] Note that it is generally used: $s_s = 1$

[b] Specific yield is determined by $\varepsilon_e = \varepsilon(1 - s_r)$

For the given case an analytical solution for the filling water height h can be easily derived as

$$h(t) = 10\left(\sqrt[3]{1 + 2.1681\,t} - 1\right) \quad (h \text{ in meters},\ t \text{ in years}) \tag{10.144}$$

The open pit reaches its maximum height with $h_o = 20$ m after $t = V_{\text{pit}}/Q =$ 12 years of filling with the given discharge Q.

To simulate the flooding process for the idealized open-pit mine both the classic free-surface modeling strategy with a moving mesh BASD technique and the unsaturated-saturated modeling strategy are applied. The used parameters and conditions are summarized in Tables 10.31 and 10.32. The simulations for both strategies are performed with the adaptive AB/TR predictor-corrector time stepping. The pit geometry solely determines the temporal storage in the flooding process. For the moving mesh BASD strategy only two layers are sufficient to describe exactly the stratigraphic relationship of the pit (Fig. 10.53a), where the upper layer represents the 'air' domain to be filled. As evidenced in Fig. 10.54 for the computed filling curve $h(t)$ the moving mesh gives an excellent agreement with the analytical solution.

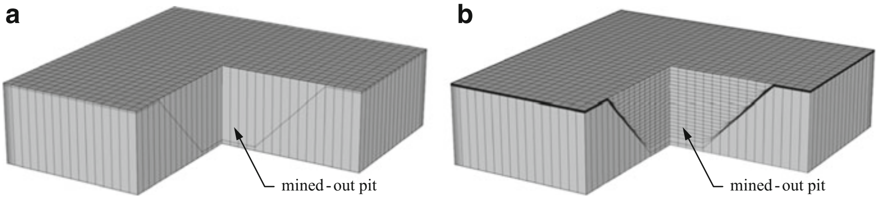


Fig. 10.53 Meshes used for the pit flooding simulation (*cut view*, vertical exaggeration 10:1). Study domain measures 800 × 800 × 21.5 m: (**a**) classic moving mesh BASD technique based on a simple two-layer 32 × 32 × 2 brick element schematization (only one layer approximates the inner mined-out pit domain) and (**b**) unsaturated-saturated model discretized with 21 layers (20 layers are used for the inner mined-out pit domain) consisting of 32 × 32 × 21 brick elements

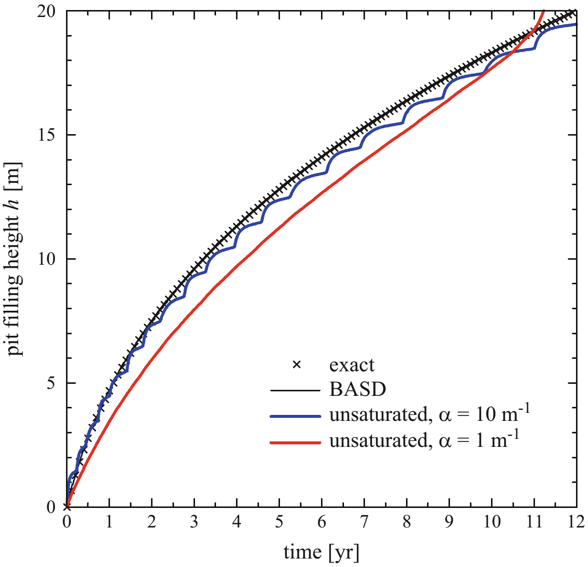


Fig. 10.54 Simulated filling curves $h(t)$. Results of the unsaturated-saturated model obtained for two different α—values in comparison to the analytic solution (10.144) and predictions from the BASD-based moving mesh strategy

While a free-surface moving mesh strategy is superior to simulate the present open-pit flooding process, a fixed mesh with an unsaturated-saturated modeling approach needs naturally more effort, however, in favor of its appropriateness in complex applications. 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' space. The unsaturated approach serves as a contrivance to smooth the numerical solutions. For the unsaturated-saturated model 21 layers are chosen to schematize the pit body as shown in Fig. 10.53b. The standard $h-$based form of the Richards' equation appears sufficient and appropriate for this type of problem. The modified van Genuchten parametric model is preferred with parameters listed in Table 10.32. The unsaturated parameters are suitably selected. Two different $\alpha-$values are tested to indicate their influence on the accuracy of the predicted filling curve. The results are shown in Fig. 10.54 in comparison to the analytical solution and the BASD-based moving mesh result. Expectedly, the smoother the capillary pressure relationship in the pit is assigned, the more the filling curve is smoothly shaped and, however, the more the rise of the water table lags. As revealed in Fig. 10.54 a reasonable accuracy is achieved for a α of $10\,\mathrm{m}^{-1}$, regardless of the somewhat wavy behavior of the filling curve. A compromise between vertical resolution and strength of the capillary pressure relationship in the pit is to be found to get acceptable results even in open-pit flooding processes.

Chapter 11
Variable-Density Flow, Mass and Heat Transport in Porous Media

11.1 Introduction

In Chap. 3 the continuum approach of the porous medium has been described. A fluid (or better a phase) appears there as an effectively continuous medium with a mass *density* ρ (fluid mass per unit volume of fluid) as a fundamental bulk property. The density of a fluid is often not uniform. In general, the fluid is composed of N miscible chemical species with a partial density ρ_k (mass of the constituent k per unit volume of fluid), so that for the mixture $\rho = \sum_k^N \rho_k$ (density increases when dissolved mass of constituents increases). Moreover, the density of a fluid can be influenced by the temperature T (density decreases when temperature increases) and by the pressure p (density increases when pressure increases due to compressibility). In a formal manner, the density is to be regarded as a dependent thermodynamic variable for which an equation of state (EOS) $\rho = \rho(p, \rho_k, T)$ holds, cf. Sect. 3.8.6.1.

Among the state variables, density merits special attention as its spatial and temporal variations are fundamental to the class of *variable-density flow*, sometimes categorized as *density-driven flow* or *buoyancy-driven flow*. Mathematically, this is expressed by the presence of ρ in the gravity (buoyancy) term $\rho \boldsymbol{g}$ appearing in the momentum balance equation for a fluid. In systems with variable density many different, yet physically correct, flow patterns may occur.[1]

[1] We note that the impact of p, ρ_k and T on ρ does not lead to the same flow effects. Compression effects caused by pressure changes will not feature a new physical characteristic, quite contrary to variable concentration or/and temperature fields which are governed by distinct balance statements subjected to advection and dispersion/conduction. Only the presence of at least one of these quantities is capable of forming complex convective flow phenomena such as flow recirculations, stratified and physically oscillating flow patterns. Flow processes affected exclusively by compression due to pressure changes will not belong to the distinct category of variable-density flow.

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_11,

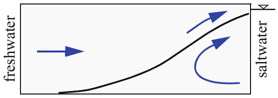
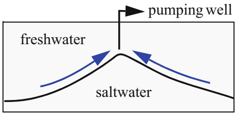
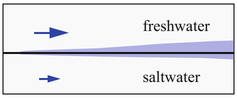
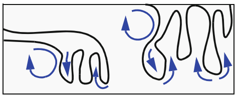
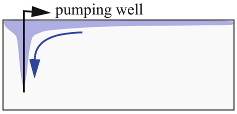
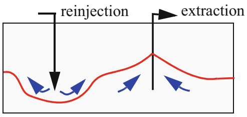
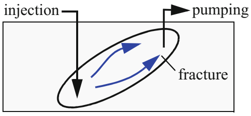
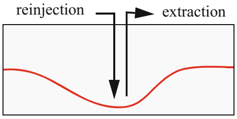
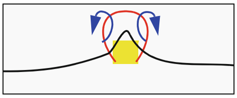
The corresponding mathematical models can imply nonunique solutions, and issues of physical stability, oscillations and chaos may arise.

In numerous natural and engineered systems, variable-density flow processes play an important role. Besides various applications in the dynamics of pure viscous fluids, atmospheric flows, oceanography, limnology, energy technology and astrophysics, we find such phenomena in many areas of subsurface hydrology, geothermics, reservoir mechanics, underground nuclear engineering and material science. Typical applications include saltwater intrusion in exploited coastal aquifers, saltwater upconing below pumping wells, concentrated brine transport, infiltration of leachates from landfills and industrial waste disposals, design of geothermal energy extraction and storage systems, large-scale convection in deep geothermal areas, radionuclides released from repository in rock salt formations and many others (Table 11.1).

Variable-density flow processes in porous media have received the attention of many researchers during the last 40 years, although, the pioneering work in this field is even older. Horton and Rogers [262], and independently Lapwood [333] first addressed the porous-medium analog of the Rayleigh-Bénard convection with regard to thermal instability in a saturated porous layer of infinite horizontal extent. Wooding [569] extended these studies, and Schneider [463] and Elder [153] performed laboratory experiments with Hele-Shaw cells. De Josselin de Jong [117] developed the vortex theory for density-driven flows in saturated porous media. While the first numerical computations of 2D convection processes in porous media were given by Wooding [569] using an iterative relaxation method, Elder [153, 154] was the first to fully compute the multicellular thermal convection currents in 2D porous layers for both steady-state and transient situations using a FDM. Since then, the number of papers on the subject of variable-density flow processes in porous media has been growing at an ever increasing rate.

Excellent reviews of prior work have been presented by Combarnous and Borries [95], Cheng [77], Gebhart et al. [187], Tien and Vafai [515], Nield and Bejan [389] and Holzbecher [255]. Most of the earlier (pre-1960) investigations were motivated by an interest in geophysical and geothermal phenomena. As subsequent studies covered an increasing range of subjects, the importance of numerical analysis soon became obvious with the application of the various numerical techniques (finite differences, Galerkin technique, spectral method, boundary element method, multigrid technique, finite elements, finite volumes, e.g., [113, 248, 252, 457, 494]). More recent reviews of modeling variable-density flow in porous media have been presented by Simmons et al. [478, 480], Diersch and Kolditz [138] and Werner et al. [561]. Among the vast work we want to highlight the following basic studies: The stability of 2D convection rolls in a porous medium heated from below was studied by Straus [494], who showed that at a given Rayleigh number less than 380, there is only a limited band of wave numbers in form of a balloon-shaped closed curve for which convective rolls are stable. Oscillatory convective currents in two dimensions were first reported by

Table 11.1 Typical applications of variable-density flow in porous (and fractured) media

Application	Case	Pattern
Saltwater	Intrusion	
	Upconing	
	Stratification/boundary layer/ flushing	
Heavy/light solutes	Fingering	
	Floating	
Geothermics	Heat extraction	
	Hot-dry-rock	
	Cooling/energy underground storage	
Underground nuclear engineering	Radionuclides released from a repository (heat sources)	
Others	Salinization of soils	
	Convection in snow layers and ice formations	
	Convection in permafrozen soils	
	Diagenetic processes in sedimentary basins	
	Drying processes in engineered and natural systems	
	Thermal isolation (pipes, dresses, hairs, rocks, wood, ...)	
	Convection in magma chambers	

Combarnous and Le Fur [96], Horne and O'Sullivan [260, 261] and simulated by Horne and Caltagirone [259] and Schubert and Straus [466] in a square cavity. Oscillatory convections in 3D porous boxes were studied by Horne [258], Schubert and Straus [465], Straus and Schubert [495, 496] and Caltagirone et al. [67]. The variation of the critical Rayleigh number (characterizing the onset of convection) and corresponding preferred cellular modes were analyzed by Beck [41] for an enclosed 3D porous medium. Numerical computations of cellular convection at high Rayleigh numbers were performed by Steen and Aidun [491], Kimura et al. [308, 309], Caltagirone et al. [68] and Caltagirone and Farbie [66]. Effects of anisotropy and heterogeneity were studied by Kvernvold and Tyvand [323], McKibbin and O'Sullivan [364, 365], McKibbin and Tyvand [366–368] and others. Stabilizing effects caused by hydrodynamic dispersion were modeled by Kvernvold and Tyvand [324] and Georgiadis and Catton [190]. Non-Darcian flow effects were considered by Katto and Masuoka [304], Walker and Homsy [556], and Prasad and Kladias [428]. Thermohaline (double-diffusive) convection processes in porous media were studied, among others, by Nield [387], Griffiths [214], Rubin [450], Rubin and Roth [451, 452], Trevisan and Bejan [518], Murray and Chen [380], Tyvand [526], Green [207], Taunton et al. [507], Goyeau et al. [197], Oldenburg and Pruess [399], Diersch and Kolditz [137], Pringle et al. [431], and Hughes et al. [271].

In hydrogeologic systems saltwater intrusion and upconing processes are a subject of specific concern. In many situations fluid-density effects are important in the vertical and horizontal displacement of saline water, which can be classified into different groups on the basis of the total dissolved-solid (TDS) concentration (3.214), see Table 11.2. Usually, the heavier saltwater underlies the lighter freshwater in a natural system and the resulting density stratification stabilizes the flow system. While the density of 'average' surface seawater ranges between 1,022 and 1,028 kg m^{-3} (typical seawater contains about 35 ppt TDS), brine densities exceeding 1,300 kg m^{-3} can occur in deep formations. The existence of high salt concentrations can give rise to large concentration gradients in the form of a narrow freshwater-saltwater transition zone. Here, the underlain salinity acts as the restoring force, while hydrodynamic dispersion and convection lead to a mixing and vertical displacement of the brine. Classically, the sharp saltwater-freshwater interface assumption is made to enable rather simple analytical and numerical solutions. This concept was used independently by Badon-Ghyben [19] and Herzberg [246] to derive a formula, today well-known as the *Ghyben-Herzberg relation*, which relates the elevation of the groundwater table to the elevation of the saltwater-freshwater interface assuming a hydrostatic equilibrium. The sharp interface approach (saltwater and freshwater as two immiscible fluids) was subsequently applied and improved in numerous works, for a review see, e.g., Reilly and Goodman [438], Bear [36], and Cheng and Quazar [78]. However, if an assessment of the salt concentration in both local and regional flow systems is desired, the more rigorous miscible-fluid approach is required. The first attempts to model the density-dependent miscible saltwater-freshwater systems applied to coastal problems were made by Henry [242] and Pinder and Cooper [420]. Due to its practical importance, the numerical modeling of saltwater intrusion and upconing

Table 11.2 Classification of salty waters

Krieger et al. [322]		Davis and De Wiest [112]	
Description	TDS (ppt)[a]	Description	TDS (ppt)[a]
Slightly saline	1−3	Freshwater	0−1
Moderately saline	3−10	Brackish water	1−10
Very saline	10−35	Saltwater	10−100
Brine	>35	Brine	>100

[a] Total dissolved solids (3.214) in parts per thousand (ppt), equivalent to ($\mathrm{g\,l^{-1}}$)

processes has received increased attention in the water resources literature over the last 30 years, resulting in better ways to model the advective and dispersive mechanisms with fluid density and viscosity effects. Saltwater intrusion processes were analyzed among others by Segol at al. [472], Segol and Pinder [471], Huyakorn and Taylor [281], Huyakorn et al. [283], Volker and Rushton [548], Frind [174], Voss [550, 551], Voss and Souza [552], Putti and Paniconi [432], Diersch [133], Galeati et al. [180], Gambolati et al. [182], Kolditz et al. [318], Bués and Oltean [63] and Abarca et al. [1]. Upconing below pumping wells was studied numerically by Diersch et al. [143], Diersch and Nillert [140], Reilly and Goodman [439] and Holzbecher [253]. Waste disposal in deep salt formations required modeling of density-dependent flow processes in the vicinity of salt domes. Studies by Herbert et al. [244], Oldenburg and Pruess [398], Oldenburg et al. [400], Johns and Rivera [290], Kolditz et al. [318], Konikow et al. [320], Holzbecher [254] and Younes et al. [583] contributed to a better informed discussion of this subject.

Variable-density flow processes in porous media are crucial in a wide spectrum of thermal and saline transport problems. Numerical studies have emphasized the importance of the Oberbeck-Boussinesq approximation and its (non-Boussinesq) extension (see Sect. 3.10.3), the physical stability and the oscillatory behavior at high density contrasts, e.g., [243, 574–576]. In the following, advanced numerical strategies for solving the coupled spatio-temporal convection processes will be considered. The consistency problem for the velocity approximation at high density variations has to be of specific interest. Various benchmark tests exist for verifying numerical models of 2D and 3D variable-density flow in porous media [138], which will be revisited and discussed.

11.2 Basic Equations

The system of the basic PDE's for 3D and 2D (including axisymmetric) variable-density flow in porous media has been developed in Sect. 3.10.5 and summarized in Table 3.7. Due to the density dependency the governing Darcy-type flow equation is nonlinearly coupled with the mass and/or heat transport equation(s) via the buoyancy term χe. We obtain the following general system of PDE's written for the convective forms of the transport equations[2]

$$\begin{aligned}
s\,S_o\frac{\partial h}{\partial t}+\varepsilon\frac{\partial s}{\partial t}+\nabla\cdot\boldsymbol{q} &= Q+Q_{\mathrm{EOB}}\\
\boldsymbol{q} &= -k_r\boldsymbol{K}f_\mu\cdot\big(\nabla h+\chi\boldsymbol{e}\big)\\
\varepsilon s\acute{\mathfrak{R}}_k\frac{\partial C_k}{\partial t}+\boldsymbol{q}\cdot\nabla C_k-\nabla\cdot(\boldsymbol{D}_k\cdot\nabla C_k)+(\varepsilon s\vartheta_k\mathfrak{R}_k+Q)C_k &= \tilde{R}_k\quad(k=1,\dots,N)\\
\big(\varepsilon s\rho c+(1-\varepsilon)\rho^s c^s\big)\frac{\partial T}{\partial t}+\rho c\boldsymbol{q}\cdot\nabla T-\nabla\cdot(\boldsymbol{\Lambda}\cdot\nabla T) &= H_e-\rho c(T-T_0)Q
\end{aligned}\tag{11.1}$$

associated with the constitutive relations

$$\begin{aligned}
\chi=\tfrac{\rho-\rho_0}{\rho_0} &= \sum_k\beta_{c_k}(C_k-C_{k0})-\beta(T)(T-T_0)\\
\beta(T) &= \begin{cases}\beta & \text{constant thermal expansion}\\ \text{Eq. (C.8)} & \text{variable thermal expansion}\end{cases}\\
\beta_{c_k} &= \tfrac{\alpha_k}{C_{ks}-C_{k0}}\\
\boldsymbol{D}_k &= \varepsilon s D_k\boldsymbol{\delta}+\boldsymbol{D}_{\mathrm{mech}}\\
\boldsymbol{\Lambda} &= \big[\varepsilon s\Lambda+(1-\varepsilon)\Lambda^s\big]\boldsymbol{\delta}+\rho c\boldsymbol{D}_{\mathrm{mech}}\\
\boldsymbol{D}_{\mathrm{mech}} &= \beta_T\|\boldsymbol{q}\|\boldsymbol{\delta}+(\beta_L-\beta_T)\tfrac{\boldsymbol{q}\otimes\boldsymbol{q}}{\|\boldsymbol{q}\|}\\
f_\mu &= \frac{\mu_0}{\mu\big(\frac{C_k}{\rho},T\big)},\ \text{e.g., Eq. (3.218)}\\
H_e &= \rho H+\rho^s H_s
\end{aligned}\tag{11.2}$$

where the remaining relations of $s=s(\psi)$, $k_r=k_r(s)$ are defined in Appendix D for variably saturated porous media, and $\mathfrak{R}_k$ is defined by (3.253) and in Table 3.8. For fully saturated porous media it is $s=k_r\equiv 1$. We note that the species indicator k runs over the number of the considered solutes (a single-species solute transport represents the special case with $k=1$, where $C\equiv C_1$ dropping the species indicator for convenience). Furthermore, notice that in (11.1) we use at first for convenience the linear Fick's law of hydrodynamic dispersion, (3.272) with $\mathfrak{I}_H=0$. The treatment of non-Fickian dispersion will be subject of Sect. 11.10. Chemically reactive multispecies processes in the fluid and solid phase will not be considered in the present chapter, they will be focused in Chap. 12. The additional sink term Q_{EOB} appearing in (11.1) is non-zero for the extended Oberbeck-Boussinesq approximation (cf. Sect. 3.10.3), given by

[2] Alternatively, by using the divergence forms of the governing transport equations for mass and heat, the coupled PDE system reads

$$\begin{aligned}
s\,S_o\tfrac{\partial h}{\partial t}+\varepsilon\tfrac{\partial s}{\partial t}+\nabla\cdot\boldsymbol{q} &= Q+Q_{\mathrm{EOB}}\\
\boldsymbol{q} &= -k_r\boldsymbol{K}f_\mu\cdot\big(\nabla h+\chi\boldsymbol{e}\big)\\
\tfrac{\partial}{\partial t}(\varepsilon s\mathfrak{R}_kC_k)+\nabla\cdot(\boldsymbol{q}C_k)-\nabla\cdot(\boldsymbol{D}_k\cdot\nabla C_k)+\varepsilon s\vartheta_k\mathfrak{R}_kC_k &= \tilde{R}_k\quad(k=1,\dots,N)\\
\tfrac{\partial}{\partial t}\big[\big(\varepsilon s\rho c+(1-\varepsilon)\rho^s c^s\big)(T-T_0)\big]+\nabla\cdot(\rho c\boldsymbol{q}(T-T_0))-\nabla\cdot(\boldsymbol{\Lambda}\cdot\nabla T) &= H_e
\end{aligned}$$

Note that the divergence form of heat transport in terms of the temperature T assumes that the specific heat capacities c and c^s are independent of T (cf. discussions in Sect. 3.9.1).

$$
\begin{aligned}
Q_{\text{EOB}} &= -\boldsymbol{q}\cdot\Big(\frac{S_o}{\varepsilon}\nabla h + \sum_k \beta_{c_k}\nabla C_k - \beta^*\nabla T\Big) - \varepsilon s\Big(\sum_k \beta_{c_k}\frac{\partial C_k}{\partial t} - \beta^*\frac{\partial T}{\partial t}\Big) \\
\beta^* &= \begin{cases} \beta & \text{constant thermal expansion} \\ \dfrac{\beta(T) + \frac{\partial\beta(T)}{\partial T}(T-T_0)}{1+\sum_k \beta_{c_k}(C_k - C_{k0}) - \beta(T)(T-T_0)} & \text{variable thermal expansion} \end{cases}
\end{aligned}
\tag{11.3}
$$

The PDE system (11.1) has to be solved for the hydraulic head h, the saturation s, the velocity $\boldsymbol{q}$, the concentration C_k of chemical species k and the temperature T. It is common practice to substitute $\boldsymbol{q}$ by the Darcy equation to obtain the governing Richards-type equation (cf. Sect. 3.11) in a form given by (10.5) and to express the saturation s via the available capillary pressure relationship (10.3) as discussed in Chap. 10. Finally, the following PDE system holds

$$
\begin{aligned}
s\,S_o\frac{\partial h}{\partial t} + \varepsilon\frac{\partial s}{\partial t} - \nabla\cdot\big[k_r\boldsymbol{K}f_\mu\cdot(\nabla h + \chi\boldsymbol{e})\big] &= Q_h + Q_{hw} + Q_{\text{EOB}} \\
\varepsilon s\acute{\mathfrak{R}}_k\frac{\partial C_k}{\partial t} + \boldsymbol{q}\cdot\nabla C_k - \nabla\cdot(\boldsymbol{D}_k\cdot\nabla C_k) + (\varepsilon s\vartheta_k\mathfrak{R}_k + Q_h)C_k &= \hat{R}_k + Q_{kw} + Q_k \;(k=1,\ldots,N) \\
\big(\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s\big)\frac{\partial T}{\partial t} + \rho c\boldsymbol{q}\cdot\nabla T - \nabla\cdot(\boldsymbol{\varLambda}\cdot\nabla T) &= Q_T + Q_{Tw} - \rho c(T-T_0)Q_h
\end{aligned}
\tag{11.4}
$$

where the source/sink terms $Q = Q_h + Q_{hw}$, $\tilde{R}_k = \hat{R}_k + Q_{kw} + Q_k$ and $H_e = Q_T + Q_{Tw}$ are suitably split into the supply terms Q_h, $\hat{R}_k$, Q_k, $Q_T = \rho H^\star + \rho^s H_s^\star$ and well-type SPC terms Q_{hw}, Q_{kw}, Q_{Tw}, respectively. The three density-coupled PDE's (11.4) have to be solved for the chosen primary variables of the hydraulic head h, the concentration C_k of chemical species k and the temperature T by using the constitutive relations (11.2) and the following set of BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC as introduced in Sects. 6.3.1–6.3.3 and 6.5.5:

$$
\begin{aligned}
h &= h_D & &\text{on}\quad \varGamma_{D_h}\times t[t_0,\infty) \\
-[k_r\boldsymbol{K}f_\mu\cdot(\nabla h + \chi\boldsymbol{e})]\cdot\boldsymbol{n} &= q_h & &\text{on}\quad \varGamma_{N_h}\times t[t_0,\infty) \\
-[\boldsymbol{K}f_\mu\cdot(1+\chi)\boldsymbol{e})]\cdot\boldsymbol{n} &= q_h^\nabla & &\text{on}\quad \varGamma_{N_h}^\nabla\times t[t_0,\infty) \\
-[k_r\boldsymbol{K}f_\mu\cdot(\nabla h + \chi\boldsymbol{e})]\cdot\boldsymbol{n} &= -\varPhi_h(h_C - h) & &\text{on}\quad \varGamma_{C_h}\times t[t_0,\infty) \\
Q_{hw} &= -\textstyle\sum_w Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) & &\text{on}\quad \boldsymbol{x}_w\in\varOmega\times t[t_0,\infty) \\[1ex]
C_k &= C_{kD} & &\text{on}\quad \varGamma_{D_k}\times t[t_0,\infty) \\
-(\boldsymbol{D}\cdot\nabla C_k)\cdot\boldsymbol{n} &= q_{kC} & &\text{on}\quad \varGamma_{N_k}\times t[t_0,\infty) \\
-(\boldsymbol{D}\cdot\nabla C_k)\cdot\boldsymbol{n} &= -\varPhi_{kC}(C_{kC} - C_k) & &\text{on}\quad \varGamma_{C_k}\times t[t_0,\infty) \\
Q_{kw} &= -\textstyle\sum_w Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w)(C_{kw} - C_k) & &\text{on}\quad \boldsymbol{x}_w\in\varOmega\times t[t_0,\infty) \\[1ex]
T &= T_D & &\text{on}\quad \varGamma_{D_T}\times t[t_0,\infty) \\
-(\boldsymbol{\varLambda}\cdot\nabla T)\cdot\boldsymbol{n} &= q_T & &\text{on}\quad \varGamma_{N_T}\times t[t_0,\infty) \\
-(\boldsymbol{\varLambda}\cdot\nabla T)\cdot\boldsymbol{n} &= -\varPhi_T(T_C - T) & &\text{on}\quad \varGamma_{C_T}\times t[t_0,\infty) \\
Q_{Tw} &= -\textstyle\sum_w \rho c Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w)(T_w - T) & &\text{on}\quad \boldsymbol{x}_w\in\varOmega\times t[t_0,\infty)
\end{aligned}
\tag{11.5}
$$

in combination with the IC's of the form

$$
\left.\begin{aligned}
h(\boldsymbol{x},t_0) &= h_0(\boldsymbol{x}) \\
C_k(\boldsymbol{x},t_0) &= C_{k,0}(\boldsymbol{x}) \\
T(\boldsymbol{x},t_0) &= T_0(\boldsymbol{x})
\end{aligned}\right\}\qquad \text{in}\quad \bar{\varOmega}
\tag{11.6}
$$

where the total boundary is $\varGamma = \varGamma_{D_h} \cup \varGamma_{N_h} \cup \varGamma_{N_h}^{\nabla} \cup \varGamma_{C_h} = \varGamma_{D_k} \cup \varGamma_{N_k} \cup \varGamma_{C_k} = \varGamma_{D_T} \cup \varGamma_{N_T} \cup \varGamma_{C_T}$. Corresponding BC's of species concentration and temperature exist for the divergence forms of the mass and heat transport equations, equivalent to (8.4). Once (11.4) has been solved, the secondary variables of Darcy velocity $\boldsymbol{q}$ and saturation s can be evaluated with known h. The essential parameters required for solving (11.4) with (11.5) and (11.6) are listed in Tables I.1–I.8, I.10–I.12, I.16 and I.17 of Appendix I in accordance with the chosen problem type. *Steady-state* flow, mass and heat transport conditions occur if $\partial h/\partial t$ ($\partial s/\partial t$ for unsaturated conditions), $\partial C_k/\partial t$ and $\partial T/\partial t$ approach to zero.[3]

11.3 Sharp Interface Approximation

A common conceptual modeling strategy in saltwater intrusion is the *sharp interface approximation*, e.g., [36, 38], where it is assumed that freshwater and saltwater are two immiscible liquids. If we neglect any hydrodispersive mixing and capillary pressure effects both liquids with their different densities and viscosities occupy the own distinct portion of the flow domain and become separated from each other by a sharp, possibly moving, interface, which has to be determined. Let us denote the freshwater and saltwater by their 'phase' superscripts fw and sw, respectively, we can deduce the governing flow equations from above for isothermal and saturated conditions as

$$\begin{aligned} S_o^{fw}\frac{\partial h^{fw}}{\partial t} - \nabla\cdot\big(\boldsymbol{K}^{fw}\cdot\nabla h^{fw}\big) &= Q^{fw} \\ S_o^{sw}\frac{\partial h^{sw}}{\partial t} - \nabla\cdot\big(\boldsymbol{K}^{sw}\cdot\nabla h^{sw}\big) &= Q^{sw} \end{aligned} \tag{11.7}$$

where the piezometric (hydraulic) heads, cf. (3.260), for freshwater and saltwater are introduced, respectively, as

$$h^{fw} = \frac{p^{fw}}{\rho^{fw} g} + z, \qquad h^{sw} = \frac{p^{sw}}{\rho^{sw} g} + z \tag{11.8}$$

Since the pressure must be equal at the interface between freshwater and saltwater, i.e.,

$$p^{fw} = p^{sw} \quad \text{on the interface} \tag{11.9}$$

[3] Optionally, FEFLOW suppresses all time derivative terms $\partial h/\partial t$, $\partial s/\partial t$, $\partial C_k/\partial t$ and $\partial T/\partial t$ for solving steady-state solutions. A specific option exists, named *steady flow – transient transport*, in which only the derivative terms of the flow equation $\partial h/\partial t$ and $\partial s/\partial t$ are dropped to exclude flow storage effects in the variable-density flow simulation. Note, however, due to the nonlinearity in the flow equation the solution of the flow must be updated at each time t once the concentration C_k and/or the temperature T change. As the result, h, s and $\boldsymbol{q}$ remain time-dependent.

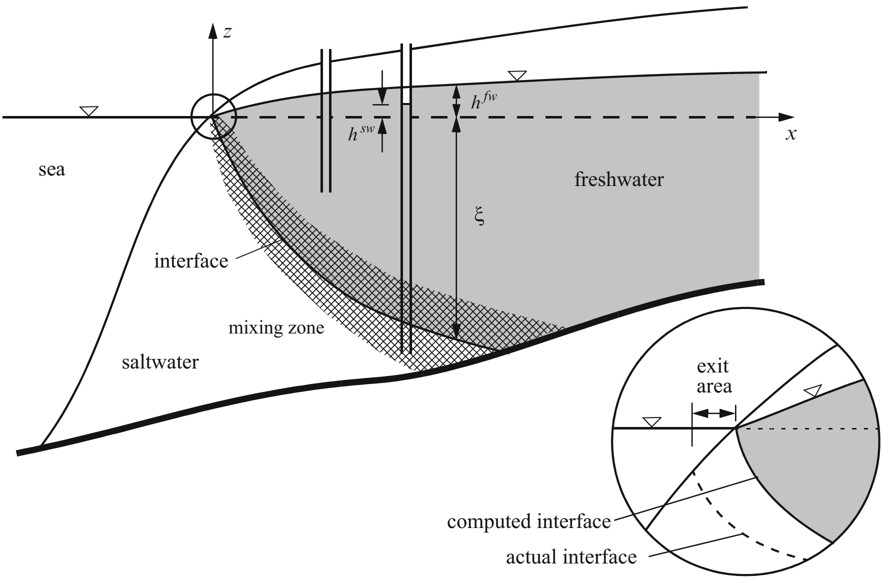


Fig. 11.1 Saltwater-freshwater interface in an unconfined coastal aquifer

we obtain from (11.8) the sought relation for the position of the sharp (possibly moving) interface in the form

$$\xi = \delta h^{fw} - (1 + \delta)h^{sw} \tag{11.10}$$

in which $\xi = -z$ represents the depth of the interface below a chosen reference datum, for instance the seawater level (see Fig. 11.1) and

$$\delta = \frac{\rho^{fw}}{\rho^{sw} - \rho^{fw}} \tag{11.11}$$

is the density ratio, assuming $\rho^{sw} > \rho^{fw}$.

The piezometric heads h^{fw} and h^{sw} have to be determined from the solution of (11.10). The interface relation (11.10), firstly introduced by Muskat [381] and Hubbert [265], can also be used to predict the interface depth ξ from observed water levels if two wells are located near to each other. Such a case is shown in Fig. 11.1, where the deeper well measures the saltwater head h^{sw} and the other shallow well measures the freshwater head h^{fw}. Then, (11.10) provides a much better estimate of ξ than the Ghyben-Herzberg relation [19,36,246], which is often practiced. Ghyben and Herzberg assumed in addition that the saltwater is *stagnant*. The pressure in the saltwater zone becomes hydostatic, i.e., $h^{sw} = \text{const}$. Using the datum shown in Fig. 11.1 it is $h^{sw} = 0$ and (11.10) reduces to the well-know *Ghyben-Herzberg relation* written as

$$\xi = \delta h^{fw} \tag{11.12}$$

The advantage of the Ghyben-Herzberg relation (11.12) is that only the water level in freshwater wells is necessary to predict the location of the stationary saltwater-freshwater interface. It means that at any distance from the sea, the saltwater-freshwater interface below sea level is δ times the height of the freshwater table above it. For example, taking an average seawater concentration of 35 ppt TDS we can estimate from (3.276) a saltwater density of about $\rho^{sw} = 1{,}025\,\text{kg}\,\text{m}^{-3}$ so that with $\rho^{fw} = 1{,}000\,\text{kg}\,\text{m}^{-3}$ we obtain $\delta = 40$. It results in this case that the stationary saltwater-freshwater below the sea level interface is 40 times the height of the freshwater table above it.

Sharp interface models do not consider hydrodispersive mixing of saltwater and freshwater, which restricts significantly their applicability in real applications. Only in cases, where the mixing zone is very narrow in relation to the vertical and horizontal extent of the study domain, they can be useful provided only the location of the saltwater-freshwater interface is of major interest. However, sharp interface models are inappropriate to predict mixing concentrations in wells or bodies of water threatened by upconing and/or spreading saltwater. Due to these reasons, today's modeling approach generally prefers the solution of the complete set of PDE's (11.4) providing a general physical description without essential restrictions and simplifications, however, at the expense of a usually higher computational effort.

11.4 Hydrostatic Condition and Evaluation of Observation Wells

A hydrostatic condition is given when either the fluid is at rest or a steady uniform flow exists, i.e., $\boldsymbol{q} = \text{const}$. Assuming that the gravity is directed vertically along the z−coordinate so that $\boldsymbol{g}^T = (0\ 0\ -g)$ and $\boldsymbol{e}^T = (0\ 0\ 1)$, it follows directly from the governing Darcy equation, (3.258), (3.262) or (11.1), that

$$\nabla_z p + \rho g = \text{const}, \qquad \nabla_z h + \chi = \text{const} \tag{11.13}$$

where the buoyancy coefficient (3.265) and the hydraulic head (3.260), respectively, are defined as

$$\chi = \frac{\rho - \rho_0}{\rho_0}, \qquad h = \frac{p}{\rho_0 g} + z \tag{11.14}$$

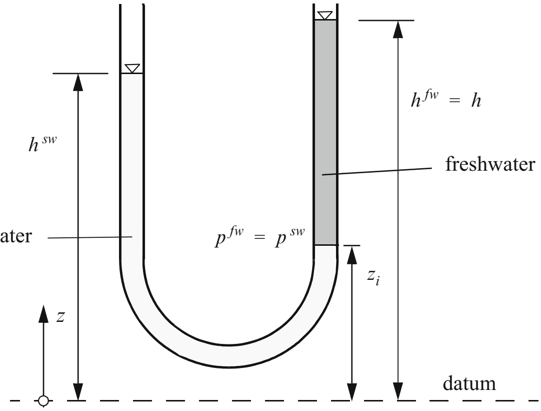


Fig. 11.2 Hydrostatic saltwater-freshwater equilibrium in a U-pipe

which are related to the constant reference liquid density ρ_0, conveniently be chosen as the freshwater density, i.e., $\rho_0 \equiv \rho^{fw}$ and $h \equiv h^{fw}$. Two situations are here of more practical interest:

(i) If we measure a saltwater head h^{sw} in a piezometric pipe, where the complete height in the pipe is filled by water of constant density ρ^{sw}, the question arises of how much is the equivalent freshwater level $h = h^{fw}$ in a piezometric pipe. By using (11.13) the pressure $p = -g \int \rho dz + C$ at the position $z = z_i$ of a saltwater-freshwater interface (Fig. 11.2) must be the same for saltwater and freshwater, viz.,

$$p(z_i) = p^{sw} = -g \int_{z_i}^{h^{sw}} \rho^{sw} dz + C \equiv p^{fw} = -g \int_{z_i}^{h^{fw}} \rho^{fw} dz + C \quad (11.15)$$

where C is an arbitrary constant ($C = 0$ for water at rest). If ρ^{sw} and ρ^{fw} are constant over the considered heights, the equilibrium condition (11.15) yields the following relationship

$$h = h^{fw} = \frac{\rho^{sw}}{\rho^{fw}} h^{sw} - \left(\frac{\rho^{sw} - \rho^{fw}}{\rho^{fw}} \right) z_i = \left(1 + \frac{1}{\delta}\right) h^{sw} - \frac{z_i}{\delta} \quad (11.16)$$

where δ is defined by (11.11). Formula (11.16) is useful to convert measured saltwater heads h^{sw} at $z = z_i$ into equivalent freshwater heads h. For example, if we measure $h^{sw} = 80$ m at elevation $z_i = 20$ m, a freshwater head of $h = 81.5$ m results with $\delta = 40$.

(ii) In practice, concentration and/or temperature-dependent density variations along a piezometric pipe can occur due to a leaky or insufficiently insulated observation well. Provided that the concentration and/or temperature distributions along the pipe are known from measurements (e.g., via conductimeters),

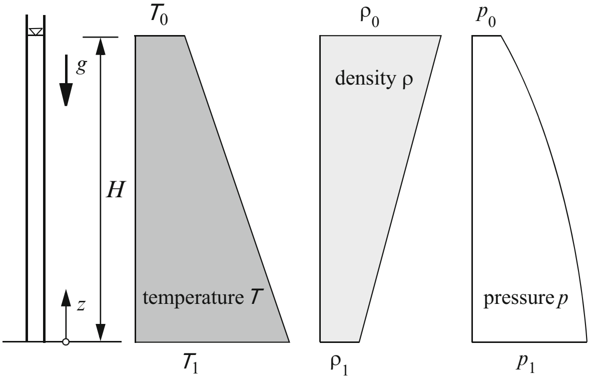


Fig. 11.3 Hydrostatic condition in a piezometric pipe with linearly varying temperature

the hydrostatic pressure at elevation z can be easily determined from

$$p(z) = -g \int_{z_1}^{z} \rho(z) dz + C \tag{11.17}$$

For example, if the temperature increases linearly with depth of a borehole $T(z) = T_1 - (T_1 - T_0)z/H$ (see Fig. 11.3), the density $\rho(z) = \rho_0[1 - \beta(T(z) - T_0)]$ decreases linearly in depth according to $\rho(z) = \rho_0[1 - \beta(T_1 - T_0)(1 - z/H)]$ and a quadratic variation for the pressure in depth results

$$p(z) = p_1 - \rho_0 g\big[z + \tfrac{\beta}{2}(T_1 - T_0)(\tfrac{z^2}{H} - 2z)\big] \tag{11.18}$$

assuming a constant thermal expansion coefficient β. Similar expressions result for a concentration profile increasing with depth of a borehole, however, providing a reverse relation.

11.5 Convection Phenomena

11.5.1 Horton-Rogers-Lapwood Problem

The Horton-Rogers-Lapwood (HRL) problem is the porous-medium analog of the Rayleigh-Bénard cellular convection problem, which was first analyzed by Horton and Rogers [262] and independently by Lapwood [333]. It refers to an infinite horizontal porous layer which is uniformly heated from below and, in addition, subjected to concentration gradient(s) [387] (see Fig. 11.4). In the original HRL problem formulation the porous medium is assumed homogeneous, isotropic and fully saturated. It is further assumed that the Darcy law and the OB approximation are valid. It is supposed that the fluid density ρ, (3.274), is a function of concentra-

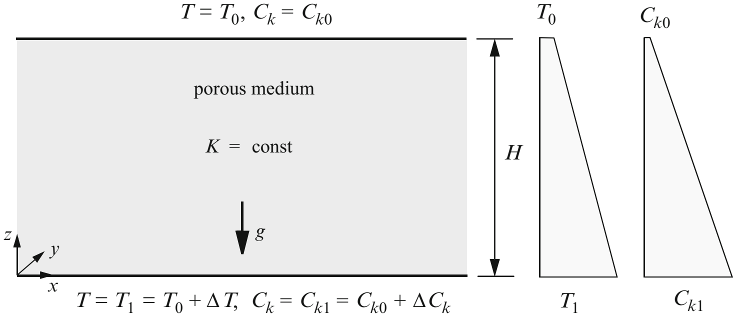


Fig. 11.4 Definition sketch for the HRL problem: infinite horizontal porous layer with linear distribution of temperature and concentration(s) where $\Delta T = T_1 - T_0$ and $\Delta C_k = C_{k1} - C_{k0}$

tion(s) C_k of non-reactive species k and temperature T written as

$$\rho = \rho_0 \big[1 + \sum_k \beta_{c_k}(C_k - C_{k0}) - \beta(T - T_0)\big] \tag{11.19}$$

where β_{c_k} and β are considered as constant solutal and thermal expansion coefficients, respectively. For the HRL problem the general balance equations (11.1) reduce to

$$\begin{aligned} \tfrac{S_o}{\varepsilon}\tfrac{\partial h}{\partial t} + \nabla \cdot \boldsymbol{v} &= 0 \\ \boldsymbol{v} &= -\tfrac{K}{\varepsilon}(\nabla h + \chi \boldsymbol{e}) \\ \tfrac{\partial C_k}{\partial t} + \boldsymbol{v} \cdot \nabla C_k - D_k \nabla^2 C_k &= 0 \quad (k = 1, \ldots, N) \\ S_\lambda \tfrac{\partial T}{\partial t} + \boldsymbol{v} \cdot \nabla T - D_\lambda \nabla^2 T &= 0 \end{aligned} \tag{11.20}$$

with the definitions

$$\begin{aligned} \boldsymbol{v} = \frac{\boldsymbol{q}}{\varepsilon}, \quad \chi &= \sum_k \beta_{c_k}(C_k - C_{k0}) - \beta(T - T_0), \\ S_\lambda = 1 + \frac{(1-\varepsilon)\rho^s c^s}{\varepsilon \rho c}, \quad D_\lambda &= \frac{\varepsilon \Lambda + (1-\varepsilon)\Lambda^s}{\varepsilon \rho c} \end{aligned} \tag{11.21}$$

where $\boldsymbol{v}$ is the intrinsic (pore) velocity, $S_\lambda \geq 1$ is the thermal storage coefficient (can also be understood as a thermal retardation factor) and D_λ is the thermal diffusivity. It is obvious that the PDE's set (11.20) has a basic steady-state solution, which satisfies the BC's $T = T_0 + \Delta T$ and $C_k = C_{k0} + \Delta C_k$ at $z = 0$ and $T = T_0$ and $C_k = C_{k0}$ at $z = H$, given by

$$\begin{aligned} \boldsymbol{q} &= \mathbf{0} \\ T &= T_1 - \Delta T \tfrac{z}{H} \\ C_k &= C_{k1} - \Delta C_k \tfrac{z}{H} \\ p &= p_1 - \rho_0 g\big[z - \tfrac{1}{2}(\beta_{c_k}\Delta C_k - \beta\Delta T)(\tfrac{z^2}{H} - 2z)\big] \\ h &= h_1 + \tfrac{1}{2}(\beta_{c_k}\Delta C_k - \beta\Delta T)(\tfrac{z^2}{H} - 2z) \end{aligned} \tag{11.22}$$

where $\Delta T = T_1 - T_0$ and $\Delta C_k = C_{k1} - C_{k0}$. This solution corresponds to the ‘diffusion state’, in which mass and heat transfer are solely by diffusion and thermal conduction, respectively.

11.5.2 Dimensionless Equations and Characteristic Numbers

To assess the relative importance of terms in the governing PDE’s set (11.20) let us introduce dimensionless variables by choosing H, $S_\lambda H^2/D_\lambda$, D_λ/H, ΔT and $\Delta C_k D_\lambda/D_k$ as scales for length, time, velocity, temperature and concentration, respectively. Thus, the dimensionless variables are defined as

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}}{H}, \quad \hat{t} = \frac{D_\lambda t}{S_\lambda H^2}, \quad \hat{\boldsymbol{v}} = \frac{H\boldsymbol{v}}{D_\lambda}, \quad \hat{T} = \frac{T}{\Delta T}, \quad \hat{C}_k = \frac{C_k}{\Delta C_k}\frac{D_k}{D_\lambda} \tag{11.23}$$

Using (11.23) the basic equations (11.20) can be written in dimensionless variables:

$$\begin{aligned} &\Big(\frac{S_o H}{\varepsilon S_\lambda}\Big)\frac{\partial \hat{h}}{\partial \hat{t}} + \hat{\nabla}\cdot\hat{\boldsymbol{v}} = 0 \\ &\hat{\boldsymbol{v}} = -\mathrm{Pe}_t\ \hat{\nabla}\hat{h} - \Big(\sum_k \mathrm{Ra}_k(\hat{C}_k - \hat{C}_{k0}) - \mathrm{Ra}_t(\hat{T} - \hat{T}_0)\Big)\boldsymbol{e} \\ &\Big(\frac{\mathrm{Le}_k}{S_\lambda}\Big)\frac{\partial \hat{C}_k}{\partial \hat{t}} + \mathrm{Le}_k\ \hat{\boldsymbol{v}}\cdot\hat{\nabla}\hat{C}_k - \hat{\nabla}^2\hat{C}_k = 0 \quad (k = 1, 2, \ldots) \\ &\frac{\partial \hat{T}}{\partial \hat{t}} + \hat{\boldsymbol{v}}\cdot\hat{\nabla}\hat{T} - \hat{\nabla}^2\hat{T} = 0 \end{aligned} \tag{11.24}$$

where $\hat{\nabla}$ is the dimensionless gradient vector and the following characteristic numbers naturally appear, viz.,

solutal Rayleigh number of species k:

$$\mathrm{Ra}_k = \frac{\beta_{c_k}\Delta C_k K H}{\varepsilon D_k} = \frac{\alpha_k K H}{\varepsilon D_k} \tag{11.25}$$

thermal Rayleigh number:

$$\mathrm{Ra}_t = \frac{\beta \Delta T K H}{\varepsilon D_\lambda} \tag{11.26}$$

thermal Darcy-Péclet number:

$$\mathrm{Pe}_t = \frac{KH}{\varepsilon D_\lambda} \tag{11.27}$$

Lewis number of species k:

$$\mathrm{Le}_k = \frac{D_\lambda}{D_k} \tag{11.28}$$

Turner (or buoyancy) number of species k:

$$\mathrm{Tu}_k = \frac{\mathrm{Ra}_k}{\mathrm{Le}_k \mathrm{Ra}_t} = \frac{\beta_{C_k} \Delta C_k}{\beta \Delta T} = \frac{\alpha_k}{\beta \Delta T} \tag{11.29}$$

The Rayleigh number relates buoyant forces to (thermal or solutal) diffusivity. It can have a positive or negative sign in dependence on the definition of ΔT or ΔC_k (cf. Fig. 11.4). Rayleigh number of zero means that density effects on the flow are excluded. The Turner number plays an important role in coupled convective mass and heat transfer, such as double-diffusive convection (DDC), which relates the buoyant forces imposed by solute and temperature differences to each other. Turner number of about unity means that solutal and thermal buoyant forces are in the same order, which induces a highly nonlinear dependency of the convection process. The Lewis number, which is often larger than unity, defines the ratio of thermal diffusivity to mass diffusivity and indicates how the diffusion speed of heat and species is related to each other. DDC phenomena are sensitively affected by the Lewis number.

In addition, to evaluate heat and mass transfer at a boundary (surface) Γ the ratio of the convective to pure conductive (diffusive) heat and mass transfer across (normal to) the boundary has relevance. For the heat and mass transfer, respectively, it is expressed in integral form by the

Nusselt number:

$$\mathrm{Nu} = \frac{\int_A q_{n_T}(t) d\Gamma}{\int_A q_{n_T}^{\mathrm{diff}} d\Gamma} \approx \frac{1}{A} \int_A (\hat{\nabla} \hat{T} \cdot \boldsymbol{n}) d\Gamma \tag{11.30}$$

Sherwood number of species k:

$$\mathrm{Sh}_k = \frac{\int_A q_{n_{kC}}(t) d\Gamma}{\int_A q_{n_{kC}}^{\mathrm{diff}} d\Gamma} \approx \frac{1}{A} \int_A (\hat{\nabla} \hat{C}_k \cdot \boldsymbol{n}) d\Gamma \tag{11.31}$$

where $q_{n_T}(t)$ and $q_{n_{kC}}(t)$ represent boundary heat and mass fluxes, respectively, which are either known from given BC's (cf. Sect. 6.3) or computable via CBFM

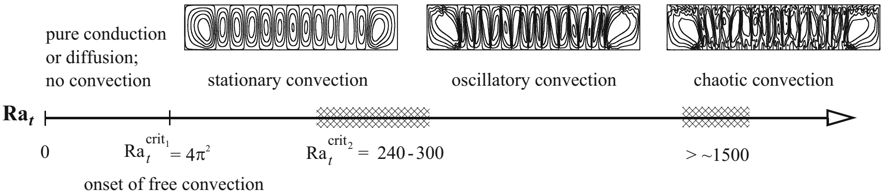


Fig. 11.5 Convection regimes of a porous layer heated from below in dependence on the thermal Rayleigh number Ra$_t$

(cf. Sect. 8.19.2), the superscript 'diff' indicates pure diffusive steady-state boundary fluxes, which are easily given for the present HRL problem as $q_{nT}^{\text{diff}} = -D_\lambda \Delta T/H$ and $q_{nkC}^{\text{diff}} = -D_k \Delta C_k/H$ (assuming $q_{nT}^{\text{diff}} \neq 0$ and $q_{nkC}^{\text{diff}} \neq 0$), and A corresponds to the exchange area.

11.5.3 Convection Regimes

11.5.3.1 Free Convection

To explain striking features of free convection let us consider next the prototypical HRL problem shown in Fig. 11.4 only for the presence of a temperature gradient (so, put aside concentration gradients at first). Starting with a temperature difference ΔT of zero, let us observe the behavior of flow and heat transfer if we continuously increase the temperature at the bottom of the porous layer, i.e., ΔT increases and the accordingly the thermal Rayleigh number Ra$_t$ starts growing from zero (Fig. 11.5). At a given Ra$_t$ the properties of the layer are considered invariable, however, it is assumed that either the velocity, the hydraulic head (equivalently the pressure) or the temperature can be anyhow perturbed at initial state by a small, possibly random finite amplitude.

1. Initially, at a small (or zero) Ra$_t$ we realize that the heat flux imposed by $\Delta T \geq 0$ can be fully processed by pure thermal conduction/diffusion, i.e., the fluid remains at rest, no convection occurs and any perturbation disappears (sooner or later). The steady-state solution corresponds to (11.22). Note that pure thermal conduction is also given for all negative Ra$_t$, where with $\Delta T < 0$ a stabilizing thermal distribution exists (light hot fluid stratifies over heavy cold fluid).
2. Now, we will reach a Rayleigh number, where ΔT has been raised high enough so that pure thermal conduction/diffusion is not capable anymore of transferring the imposed heat through the porous layer. This is the moment when the fluid must start to move: the onset of free convection at a first critical Rayleigh number Ra$_t^{\text{crit}_1}$, the system becomes physically instable. This stability criterion can be analytically determined by a linear stability analysis [389]. For the HRL problem it gives Ra$_t^{\text{crit}_1} = 4\pi^2$. Different Ra$_t^{\text{crit}_1}$ exists for a 3D porous box of finite

lengths [41] showing dependencies on the geometric aspect ratio. In general, there are also influences from BC's, the presence of mechanical dispersion, anisotropic media, layered and sloped structures, e.g., [77,95,389]. To trigger this instability, perturbation (noise) is needed because a perfectly undisturbed system will not become instable. However, a real physical system has always inherent perturbations and even in numerical modeling truncation and roundoff errors are basically present and induce perturbations, intentionally or not. The movement of fluid in the layer is characterized by a specific behavior. The fluid circulates in form of rolls and exhibits cellular convective patterns. This process is fully self-organizing, an interplay between upwelling fluid driven by buoyant forces and downwelling fluid caused due to fluid mass conservation. Mathematically, the former is implied by the buoyancy term χe and the latter is controlled via ∇h in the governing (momentum) Darcy equation. For moderate Rayleigh numbers $\mathrm{Ra}_t > \mathrm{Ra}_t^{\mathrm{crit}_1}$ we observe stationary convective rolls.

3. Stationary convection exists up to a second critical Rayleigh number $\mathrm{Ra}_t^{\mathrm{crit}_2}$, which cannot be exactly determined and is only assessable from numerical computations. Straus [494] found a number of about 380, which can be considered as an upper limit of $\mathrm{Ra}_t^{\mathrm{crit}_2}$. Other numerical investigations [67,260,261] indicate a range between $4\pi^2 < \mathrm{Ra}_t <$ (240–300), where stationary convection pattern develop in form of 2D rolls rotating in clockwise or counter-clockwise direction, 3D rolls or 3D polyhedral cells. Once $\mathrm{Ra}_t > \mathrm{Ra}_t^{\mathrm{crit}_2} \approx$ (240–300), the convection can begin to oscillate. Transient fluctuations appear in form of periodic convective patterns. It allows the sytem to 'pump' the imposed heat through the layer in a larger extent. This convective regime implies sharper inherent gradients and instable boundary layers. They cause oscillations which can be interpreted as the continuous creation and disappearance of convective plumes. A larger number of irregular plume patterns evolves over time. The convection process becomes increasingly unpredictable because the inherent, usually unknown perturbations determine the convection in a virtually uncontrollable manner, even when the perturbations are basically very small. Perturbation can result in manifold solutions (bifurcations). Possibly, the convection possesses different solutions and a numerical simulation converges to only one solution. As already observed by Horne and O'Sullivan [260] there is a possibility of either a steady multicellular structure or a fluctuating unicellular structure. Once formed, these two structures are not easily interchangeable, but the system may be assisted into either mode by a suitable perturbation during its early development. Higher-order transitions have been studied in [66,68] for a 2D square porous cavity. They showed that a second bifurcation exists, occurring at $\mathrm{Ra}_t = 390$. At this Rayleigh number the flow becomes periodic. Between 390 and 600 the process is single-periodic. Increasing Ra_t further, the flow is again periodic up to $\mathrm{Ra}_t = 1{,}000$. A quasi-periodic regime can maintain up to $\mathrm{Ra}_t = 1{,}500$, after which the single convecting roll splits into two unsteady cells by entering a chaotic restructuring (i.e., fluctuating) regime. Techniques of bifurcation theory have been used in two dimensions, among others, by Riley and Winters [444], Vadasz [530] and Vadasz and Olek [531–533]. Vadasz and Olek [533] pointed out the significance

of including a time derivative term in the momentum (DBF-type) equation, cf. (3.232), when studying wave (oscillatory) phenomena.

There is a direct analogy of the above thermal free convection for solute free convection if a concentration gradient is imposed on the porous layer (Fig. 11.4) under isothermal condition. However, solute-driven free convection can only occur if the concentration gradient acts destabilizing, i.e., a higher concentration must be on top of the layer so that $C_{k0} > C_{k1}$ and $\Delta C_k = C_{k1} - C_{k0}$ is negative. We can observe the same convection regimes for growing solutal Rayleigh numbers $|\mathrm{Ra}_k|$ analogously to Ra_t. Physically, however, we note that $|\Delta C_k|$ has an upper limit due to a maximally dissolvable mass of species and, thus, a high $|\mathrm{Ra}_k|$ can usually only be associated with a large hydraulic conductivity K, large layer thickness H, small porosity ε and/or small diffusivity D_k.

For sufficiently high Rayleigh numbers the flow regime can become physically unstable. This is triggered and controlled by perturbations. Such perturbations can have true physical meaning or can be purely numerical. It becomes clear, that accuracy and stability of the numerical solution approach are essential. The conflict that arises from a certain mathematical solution of such a class of problems was already indicated by Horne and Caltagirone [259] who concluded: '*This nonlinear problem has a plethora of possible alternative flow regimes and histories depending on the conditions applied initially and subsequently. Therefore the too-perfect conditions that are achieved using analytical or numerical techniques (paradoxically the most accurate ones in particular) may give rise to other artificial solutions that are divorced from the flow observed in noisy physical systems. … It is perhaps time to admit that mathematical solutions to nonlinear problems must of necessity include non-deterministic forcing effects in order to avoid solutions mathematically correct but physically unlikely*'. The more it is important, first, to take into account all relevant processes that occur in the physical system and, second, to fully explore the structure of the solution to the mathematical model, which enables the possible states of the system to be determined.

11.5.3.2 Double-Diffusive (Thermohaline) Convection

Double-diffusive convection (DDC) is a fundamental fluid dynamic process [522, 523], for instance responsible for large-scale circulation in oceans [53], and can also be recognized in porous-media problems [389]. It represents a buoyancy-driven transport process, which is simultaneously coupled by more than one diffusing property (chemical substances, thermal energy). For the DDC phenomenon to occur, the following three conditions must be met: (1) there should be a vertical gradient in two or more properties affecting the fluid density (e.g., concentrations of chemical species, temperature), (2) the resulting gradients in the fluid density must have opposing signs, and (3) the diffusivities of the properties must be different. In Fig. 11.6, ρ_1 and ρ_2 denote the distribution of density components to the two

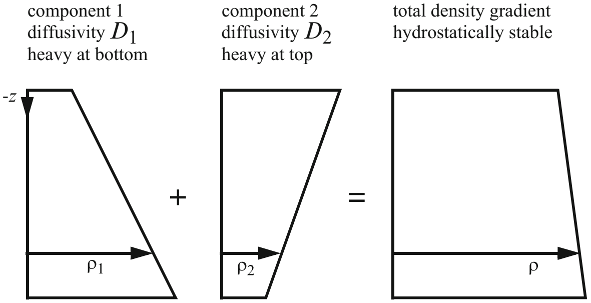


Fig. 11.6 Two density components ρ_1 and ρ_2 (with different diffusivities $D_1 \neq D_2$) and total density ρ as a function of depth (Modified from [524])

different properties mentioned above, and $\rho = \rho_1 + \rho_2$ denotes the total density distribution, cf. (2.123) and (3.194).

A striking and most surprising feature of DDC is that physical instabilities can arise even when the total density is increasing downwards, i.e., in a hydrostatically stable fluid. An important subclass of DDC phenomena represents the so-called *thermohaline flows* where the two stratifying properties consist of heat and salt. Here, heat is usually associated with the larger diffusivity value. For subsurface thermohaline processes with their two (top/bottom) configurations and two (heat/salt) properties, there are $2^2 = 4$ combinations as shown in Fig. 11.7 characterizing the four major thermohaline regimes: (a) hot and salty below (HSB), (b) cold and salty below (CSB), (c) hot and salty above (HSA) and (d) cold and salty above (CSA). The physical stability of these configurations was studied by Nield [387, 388] for the HRL problem which was generalized to a thermohaline flow (Fig. 11.4). He could show that within the $(\mathrm{Ra}_t, \mathrm{Ra}_k)$ number space there are stable and instable regions (see Fig. 11.8). We recognize two bounding lines: (1) The line $\mathrm{Ra}_t - \mathrm{Ra}_k = 4\pi^2$ represents the boundary between stable and unstable monotonic convection. (2) Additionally, a regime of oscillatory convection can be identified which is lower bounded by the line $\Phi_k \mathrm{Ra}_t - \mathrm{Ra}_k = 4\pi^2(1 + \Phi_k)$, where $\Phi_k = \mathrm{Le}_k / S_\lambda$ is a ratio of diffusivities of heat and species k formed with the thermal storage coefficient S_λ as defined in (11.21). If $\Phi_k = 1$, then both lines are parallel. Otherwise, they intersect at $\mathrm{Ra}_t = 4\pi^2\Phi_k/(\Phi_k - 1)$ and $\mathrm{Ra}_k = 4\pi^2/(\Phi_k - 1)$. Figure 11.8 illustrates the normal case with $\Phi_k > 1$, which corresponds to a Lewis number $\mathrm{Le}_k > S_\lambda \geq 1$. The combination of Ra_t and Ra_k characterizes the different DDC regimes which are of particular interest as follows:

Stable convection:

It represents a diffusive regime in which no convective currents occur. It is valid for sufficiently large Ra_k and small/negative Ra_t: $\mathrm{Ra}_t - \mathrm{Ra}_k < 4\pi^2$ and $\Phi_k \mathrm{Ra}_t - \mathrm{Ra}_k < 4\pi^2(1 + \Phi_k)$ (see Fig. 11.8). A CSB configuration (Fig. 11.7) gives always a diffusive regime.

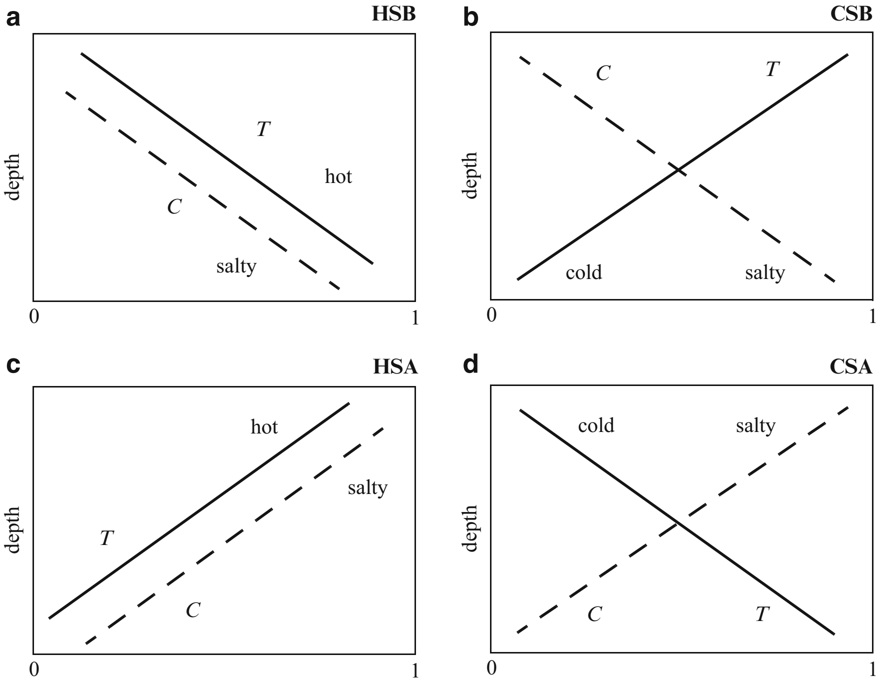


Fig. 11.7 Depth profiles of temperature T and salt C for different thermohaline regimes: (**a**) HSB (hot and salty below), heat is destabilizing, salt is stabilizing, $Q1$ stability quadrant of Fig. 11.8 (stable, monotonic or oscillatory convection), (**b**) cold and salty below (CSB), completely stabilized situation, $Q2$ stability quadrant of Fig. 11.8 (always stable diffusive regime), (**c**) hot and salty above (HSA), heat is stabilizing, salt is destabilizing, $Q3$ stability quadrant of Fig. 11.8 (stable or monotonic convection), (**d**) cold and salty above (CSA), both components acting in destabilizing manner, $Q4$ stability quadrant of Fig. 11.8 (stable or monotonic convection) (Modified from [223])

Monotonic convection:

It is also referred to as the *fingering regime* and occurs when the difference of the solutal and thermal Rayleigh numbers exceeds a critical Rayleigh number: $\mathrm{Ra}_t - \mathrm{Ra}_k \geq 4\pi^2$ (Fig. 11.8). It can potentially occur in all three configurations HSB, HSA and CSA. This regime is also termed *supercritical*. Supercriticality means that the driving (destabilizing) force exceeds the restoring (stabilizing) force. For instance, a destabilizing heavy saline fluid in a HSA configuration overcomes the stabilizing influence of heat. Usually, the thermal diffusivity is distinctly larger than solute diffusivity. If a parcel of hot salty fluid becomes perturbed downward it cools with the surround, whereby it becomes heavier due to thermal diffusivity which acts more rapidly than its dilution via the slower solute diffusivity so that the parcel continues to fall. Likewise, fluid parcels perturbed upward continue to rise. This mode of convection results in long narrow lobes of descending and rising fluid in form of *fingering* patterns, termed as *double-diffusive finger convection* (DDFC). In this regime the mass and heat transport occurs much

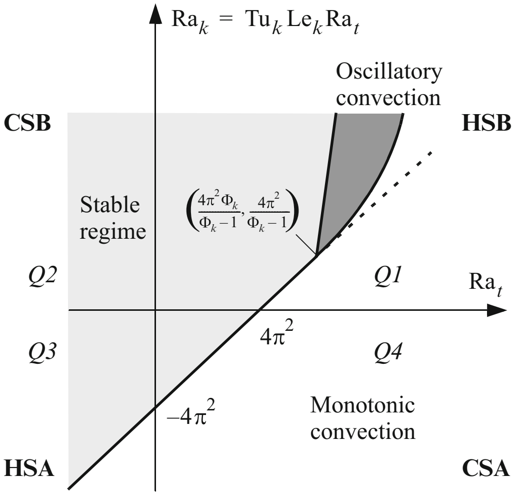

$\mathrm{Ra}_k = \mathrm{Tu}_k \mathrm{Le}_k \mathrm{Ra}_t$
Oscillatory convection
CSB
HSB
Stable regime
$\left(\frac{4\pi^2 \Phi_k}{\Phi_k - 1}, \frac{4\pi^2}{\Phi_k - 1}\right)$
Q2
Q1
Ra_t
Q3
$4\pi^2$
Q4
Monotonic convection
$-4\pi^2$
HSA
CSA

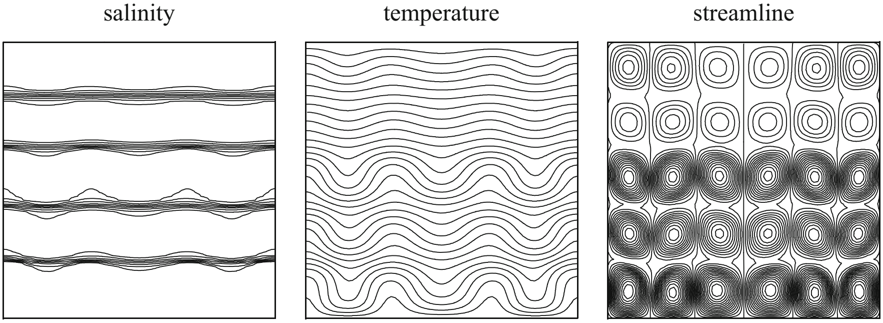
Fig. 11.8 Schematic diagram in (Ra_t, Ra_k) number space showing the stability regimes for DDC of the thermohaline HRL problem of Fig. 11.4. The *line* $\mathrm{Ra}_t - \mathrm{Ra}_k = 4\pi^2$ represents the boundary between stable and monotonic convection. The four different quadrants represent different thermohaline regimes: $Q1$, HSB (hot and salty below); $Q2$, CSB (cold and salty below); $Q3$, HSA (hot and salty above); $Q4$, CSA (cold and salty above) (Modified from [389])

faster than would be predicted by pure diffusion/conduction alone. An example of DDFC is analyzed in Sect. 11.11.8.

Oscillatory convection:

It can only occur for a HSB configuration if $\Phi_k \mathrm{Ra}_t - \mathrm{Ra}_k \geq 4\pi^2(1 + \Phi_k)$ with $\Phi_k > 1$ (i.e., $\mathrm{Le}_k > S_\lambda \geq 1$), $\mathrm{Ra}_k > 4\pi^2/(\Phi_k - 1)$ and $\mathrm{Ra}_t > 4\pi^2 \Phi_k(\Phi_k - 1)$ (cf. Figs. 11.7 and 11.8), where cold non-salty fluid overlies hot and salty fluid (in terms of a thermohaline convection). The faster diffusing heat is the destabilizing component while the slower diffusing salt is stabilizing. This regime is also termed *subcritical* because it can take place under even statically stable circumstances. Indeed, its fascinating feature is that instabilities in a basically stable system can occur due to the phase lag from the different diffusivities by heat and salt. The oscillatory mode is driven as follows: Considering a parcel of hot salty fluid which is perturbed upward, it will diffuse heat more rapidly than it diffuses salt when rising. Since it loses heat more rapidly than it loses salt, the parcel eventually becomes heavier than the surrounding fluid at which point it must begin to descend. Now, the fluid parcel will descend beyond its original position, warming as it sinks. At the new lower position the parcel eventually becomes less dense than the surrounding fluid at which point it begins again to rise and repeats the motion. A reverse circulating mechanism can be understood for fluid parcels initially perturbed downward. Finally, an oscillatory convective motion results in form of a staircase-type pattern [522,523] exhibiting well-mixed convecting layers separated by (more or less) sharp interfaces. A simulation example of a staircase DDC pattern is shown in Fig. 11.9.

The DDC phenomena discussed above for thermohaline convection can also exist in chemical systems consisting of two (or even more) species having, however, different (fast versus slow) diffusivities. An example of DDFC is described in Sect. 11.11.8 for two solutes.


salinity

temperature

streamline

Fig. 11.9 Salinity, temperature and streamline fields for a staircase situation of a thermohaline flow in a 2D square cavity at $\mathrm{Ra}_t = 400$, $\mathrm{Tu} = 2$ and $\mathrm{Le} = 100$. Salinity consists of five distinct layers with the heavy (single-species) solute on the *bottom*. The hot temperature is on the bottom too

11.6 Finite Element Formulation

Based on the principles of FEM thoroughly described in Chap. 8 we apply now the GFEM to solve the PDE system of coupled flow, mass and heat transport equations (11.4) associated with the constitutive relations (11.2), the OB extension (11.3) as well as the corresponding BC's (11.5) and IC's (11.6). For convenience we only develop the finite element equations for the convective forms of the governing mass and heat transport equations applied to fully 3D, vertical 2D and axisymmetric problems. Their alternative divergence forms will be equivalent to the formulations given in Sects. 8.5.1 and 8.9 for the general transport equation. In Sect. 11.9 the special case of variable-density problems in 2D horizontally schematized aquifers with a sloped or curved geometry will be considered.

11.6.1 Weak Forms

According to Sect. 8.5 the weak forms for the three governing PDE's (11.4) of flow, mass and heat transport can be derived analogously to the expressions (10.26) and (8.55). We obtain

$$\begin{gathered}\int_\Omega w s S_o \frac{\partial h}{\partial t} d\Omega + \int_\Omega w \varepsilon \frac{\partial s}{\partial t} d\Omega + \\ \int_\Omega \nabla w \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega - \int_\Omega w (Q_h + Q_{\mathrm{EOB}}) d\Omega + \\ \sum_w w(\boldsymbol{x}_w) Q_w(t) + \int_{\Gamma_{N_h}} w q_h d\Gamma + \int_{\Gamma_{N_h}^{\nabla}} w k_r q_h^{\nabla} d\Gamma - \\ \int_{\Gamma_{C_h}} w \Phi_h (h_C - h) d\Gamma = 0, \quad \forall w \in H_0^1(\Omega)\end{gathered} \tag{11.32}$$

$$\begin{aligned}&\int_{\Omega} w\varepsilon s\acute{\Re}_k \frac{\partial C_k}{\partial t} d\Omega + \int_{\Omega} w\boldsymbol{q}\cdot\nabla C_k d\Omega + \int_{\Omega} \nabla w\cdot(\boldsymbol{D}_k\cdot\nabla C_k)d\Omega + \\ &\int_{\Omega} w[(\varepsilon s\vartheta_k\Re_k + Q_h)C_k - \hat{R}_k - Q_k]d\Omega + \sum_w w(\boldsymbol{x}_w)Q_w(t)(C_{kw} - C_k) + \\ &\int_{\Gamma_{N_k}} wq_{kC}d\Gamma - \int_{\Gamma_{C_k}} w\Phi_{kC}(C_{kC} - C_k)d\Gamma = 0, \quad \forall w \in H_0^1(\Omega)\end{aligned} \tag{11.33}$$

$$\begin{aligned}&\int_{\Omega} w\big(\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s\big)\frac{\partial T}{\partial t}d\Omega + \int_{\Omega} w\rho c\boldsymbol{q}\cdot\nabla T d\Omega + \\ &\int_{\Omega}\nabla w\cdot(\boldsymbol{\Lambda}\cdot\nabla T)d\Omega + \int_{\Omega} w[\rho c Q_h(T - T_0) - Q_T]d\Omega + \\ &\sum_w \rho c w(\boldsymbol{x}_w)Q_w(t)(T_w - T) + \int_{\Gamma_{N_T}} wq_T d\Gamma - \\ &\int_{\Gamma_{C_T}} w\Phi_T(T_C - T)d\Gamma = 0, \quad \forall w \in H_0^1(\Omega)\end{aligned} \tag{11.34}$$

where w is a suitable weighting function and the boundary integrals are suitably separated into their segments $\Gamma = \Gamma_{D_h} \cup \Gamma_{N_h} \cup \Gamma_{N_h}^{\nabla} \cup \Gamma_{C_h} = \Gamma_{D_k} \cup \Gamma_{N_k} \cup \Gamma_{C_k} = \Gamma_{D_T} \cup \Gamma_{N_T} \cup \Gamma_{C_T}$ imposed by the Dirichlet, Neumann, gradient and Cauchy-type BC's (11.5).

11.6.2 GFEM and Resulting Nonlinear Matrix System

The weak statements (11.32)–(11.34) involve the four unknown variables h, s, C_k and T. In using the FEM these variables are replaced by a *continuous approximation* that assumes the separability of space and time (see Sect. 8.4). Thus

$$\left.\begin{aligned} h(\boldsymbol{x},t) &\approx \textstyle\sum_j N_j(\boldsymbol{x})h_j(t) \\ s(\boldsymbol{x},t) &\approx \textstyle\sum_j N_j(\boldsymbol{x})s_j(t) \\ C_k(\boldsymbol{x},t) &\approx \textstyle\sum_j N_j(\boldsymbol{x})C_{kj}(t) \\ T(\boldsymbol{x},t) &\approx \textstyle\sum_j N_j(\boldsymbol{x})T_j(t) \end{aligned}\right\} \quad \begin{aligned} j &= 1,\ldots,N_{\mathrm{P}} \\ k &= 1,\ldots,N \end{aligned} \tag{11.35}$$

where j designates global nodal indices. Using the Galerkin method with the weighting function

$$w \rightarrow w_i = N_i, \quad i = 1,\ldots,N_{\mathrm{P}} \tag{11.36}$$

and applying the approximate solutions (11.35) in (11.32)–(11.34), we obtain the following matrix systems of each N_{P} equations (cf. Sects. 8.9 and 10.5.2) written as

$$\begin{aligned} \boldsymbol{O}(\boldsymbol{h}) \cdot \dot{\boldsymbol{h}} + \boldsymbol{B} \cdot \dot{\boldsymbol{s}}(\boldsymbol{h}) + \boldsymbol{S}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) \cdot \boldsymbol{h} - \boldsymbol{F}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) &= \boldsymbol{0} \\ \boldsymbol{H}_k(\boldsymbol{h}, \boldsymbol{C}_k) \cdot \dot{\boldsymbol{C}}_k + \boldsymbol{E}_k(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) \cdot \boldsymbol{C}_k - \boldsymbol{R}_k &= \boldsymbol{0} \quad (k = 1, \ldots, N) \\ \boldsymbol{P}(\boldsymbol{h}) \cdot \dot{\boldsymbol{T}} + \boldsymbol{L}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) \cdot \boldsymbol{T} - \boldsymbol{W} &= \boldsymbol{0} \end{aligned} \tag{11.37}$$

or alternatively written in a compact form as

$$\boldsymbol{G}(\boldsymbol{U}) \cdot \dot{\boldsymbol{U}} + \boldsymbol{K}(\boldsymbol{U}) \cdot \boldsymbol{U} = \boldsymbol{Q}(\boldsymbol{U}) \tag{11.38}$$

with

$$\begin{aligned} \boldsymbol{G}(\boldsymbol{U}) &= \begin{pmatrix} \boldsymbol{O}(\boldsymbol{h}) + \boldsymbol{B} \cdot \frac{\partial \boldsymbol{s}(\boldsymbol{h})}{\partial \boldsymbol{h}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_k(\boldsymbol{h}, \boldsymbol{C}_k) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}(\boldsymbol{h}) \end{pmatrix} \\ \boldsymbol{K}(\boldsymbol{U}) &= \begin{pmatrix} \boldsymbol{S}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E}_k(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{L}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) \end{pmatrix} \\ \boldsymbol{Q}(\boldsymbol{U}) &= \begin{pmatrix} \boldsymbol{F}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) \\ \boldsymbol{R}_k \\ \boldsymbol{W} \end{pmatrix} \end{aligned} \tag{11.39}$$

and

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{h} \\ \boldsymbol{C}_k \\ \boldsymbol{T} \end{pmatrix}, \quad \dot{\boldsymbol{U}} = \begin{pmatrix} \dot{\boldsymbol{h}} \\ \dot{\boldsymbol{C}}_k \\ \dot{\boldsymbol{T}} \end{pmatrix} \tag{11.40}$$

$$\begin{aligned} &\boldsymbol{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_{N_{\mathrm{P}}} \end{pmatrix}, \quad \dot{\boldsymbol{h}} = \begin{pmatrix} \frac{dh_1}{dt} \\ \frac{dh_2}{dt} \\ \vdots \\ \frac{dh_{N_{\mathrm{P}}}}{dt} \end{pmatrix}, \quad \boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{N_{\mathrm{P}}} \end{pmatrix}, \quad \dot{\boldsymbol{s}} = \begin{pmatrix} \frac{ds_1}{dt} \\ \frac{ds_2}{dt} \\ \vdots \\ \frac{ds_{N_{\mathrm{P}}}}{dt} \end{pmatrix}, \\ &\boldsymbol{C}_k = \begin{pmatrix} C_{k1} \\ C_{k2} \\ \vdots \\ C_{kN_{\mathrm{P}}} \end{pmatrix}, \quad \dot{\boldsymbol{C}}_k = \begin{pmatrix} \frac{dC_{k1}}{dt} \\ \frac{dC_{k2}}{dt} \\ \vdots \\ \frac{dC_{kN_{\mathrm{P}}}}{dt} \end{pmatrix}, \quad \boldsymbol{T} = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_{N_{\mathrm{P}}} \end{pmatrix}, \quad \dot{\boldsymbol{T}} = \begin{pmatrix} \frac{dT_1}{dt} \\ \frac{dT_2}{dt} \\ \vdots \\ \frac{dT_{N_{\mathrm{P}}}}{dt} \end{pmatrix} \end{aligned} \tag{11.41}$$

showing the major nonlinearities in parentheses, where the matrices and RHS vectors are given as

$$\begin{aligned}
\boldsymbol{O} = O_{ij} &= \sum_e \int_{\Omega^e} s^e(h^e) S_o^e \, N_i N_j \, d\Omega^e \\
\boldsymbol{B} = B_{ij} &= \sum_e \delta_{ij} \int_{\Omega^e} \varepsilon^e \, N_i \, d\Omega^e \\
\boldsymbol{H}_k = H_{ij,k} &= \sum_e \int_{\Omega^e} \varepsilon^e s^e(h^e) \acute{\mathfrak{R}}_k^e(C_k^e) \, N_i N_j \, d\Omega^e \\
\boldsymbol{P} = P_{ij} &= \sum_e \int_{\Omega^e} \big(\varepsilon^e s^e(h^e) \rho^e c^e + (1-\varepsilon^e) \rho^{s\,e} c^{s\,e}\big) \, N_i N_j \, d\Omega^e
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{S} = S_{ij} &= \sum_e \Big(\int_{\Omega^e} \nabla N_i \cdot \big(k_r^e(s^e(h^e)) \boldsymbol{K}^e f_\mu^e(C_k^e, T^e) \cdot \nabla N_j\big) d\Omega^e + \\
&\qquad \int_{\Gamma_{C_h}^e} \Phi_h^e N_i N_j \, d\Gamma^e \Big) \\
\boldsymbol{E}_k = E_{ij,k} &= \sum_e \Big(\int_{\Omega^e} N_i \boldsymbol{q}^e(h^e, C_k^e, T^e) \cdot \nabla N_j \, d\Omega^e + \int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{D}_k^e \cdot \nabla N_j) d\Omega^e + \\
&\qquad \int_{\Omega^e} (\varepsilon^e s^e(h^e) \vartheta_k^e \mathfrak{R}_k^e + Q_h^e) N_i N_j \, d\Omega^e + \\
&\qquad \int_{\Gamma_{C_k}^e} \Phi_{kC}^e N_i N_j \, d\Gamma^e \Big) - \delta_{ij} Q_w(t)\big|_i \\
\boldsymbol{L} = L_{ij} &= \sum_e \Big(\int_{\Omega^e} N_i \rho^e c^e \boldsymbol{q}^e(h^e, C_k^e, T^e) \cdot \nabla N_j \, d\Omega^e + \int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{\Lambda}^e \cdot \nabla N_j) d\Omega^e + \\
&\qquad \int_{\Omega^e} \rho^e c^e Q_h^e N_i N_j \, d\Omega^e + \int_{\Gamma_{C_T}^e} \Phi_T^e N_i N_j \, d\Gamma^e \Big) - \delta_{ij} \rho c \, Q_w(t)\big|_i \\
\boldsymbol{F} = F_i &= \sum_e \Big(\int_{\Omega^e} N_i \big(Q_h^e + Q_{\mathrm{EOB}}^e(C_k^e, T^e)\big) d\Omega^e - \\
&\qquad \int_{\Omega^e} \nabla N_i \cdot \big(k_r^e(s^e(h^e)) \boldsymbol{K}^e f_\mu^e(C_k^e, T^e) \cdot \chi^e(C_k^e, T^e) \boldsymbol{e}\big) d\Omega^e + \\
&\qquad \int_{\Gamma_{C_h}^e} N_i \Phi_h^e h_C^e \, d\Gamma^e - \int_{\Gamma_{N_h}^e} N_i q_h^e \, d\Gamma^e - \\
&\qquad \int_{\Gamma_{N_h}^{\nabla e}} N_i k_r^e(s^e(h^e)) q_h^{\nabla e} \, d\Gamma^e \Big) - Q_w(t)\big|_i \\
\boldsymbol{R}_k = R_{i,k} &= \sum_e \Big(\int_{\Omega^e} N_i (\hat{R}_k^e + Q_k^e) d\Omega^e + \int_{\Gamma_{C_k}^e} N_i \Phi_{kC}^e C_{kC}^e \, d\Gamma^e - \\
&\qquad \int_{\Gamma_{N_k}^e} N_i q_{kC}^e \, d\Gamma^e \Big) - C_{kw} Q_w(t)\big|_i \\
\boldsymbol{W} = W_i &= \sum_e \Big(\int_{\Omega^e} N_i Q_T^e \, d\Omega^e + \int_{\Omega^e} N_i \rho^e c^e T_0^e Q_h^e \, d\Omega^e + \int_{\Gamma_{C_T}^e} N_i \Phi_T^e T_C^e \, d\Gamma^e - \\
&\qquad \int_{\Gamma_{N_T}^e} N_i q_T^e \, d\Gamma^e \Big) - \rho c T_w Q_w(t)\big|_i
\end{aligned} \tag{11.42}$$

where $(i, j = 1, \ldots, N_P)$, $(e = 1, \ldots, N_E)$ and $(k = 1, \ldots, N)$. Note that $\boldsymbol{D}_k$ and $\boldsymbol{\Lambda}$ are also functions of saturation $s = s(h)$ and Darcy velocity $\boldsymbol{q} = \boldsymbol{q}(h, C_k, T)$ and accordingly nonlinearly dependent on h, C_k and T: $\boldsymbol{D}_k = \boldsymbol{D}_k(h, C_k, T)$ and $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(h, C_k, T)$. The integrals appearing in (11.42) are integrated on element level in the local coordinates as described in Sect. 8.12. Analytical evaluations of partial integral terms of (11.42) can be deduced from developments done in Appendix H for selected element types. The differential elements $d\Omega^e$ and $d\Gamma^e$ differ for 3D, 2D and axisymmetric problems as given by (8.122)–(8.124), respectively. The tensor of the saturated hydraulic conductivity $\boldsymbol{K}^e$ of element e may be fully anisotropic in formulations introduced in Chap. 7. Is is important to note that the resulting global system of equations (11.38) is *unsymmetric* since the matrices $\boldsymbol{E}_k$ and $\boldsymbol{L}$ are unsymmetric.

The Richards-type unsaturated flow equation is discretized above in the standard $h-$based form, which is usually preferred for moderate saturation behaviors and under full saturation. However, alternative formulations exist for unsaturated flow conditions in case of need, in particular the mixed $h - s-$based form of Richards' equation by employing Celia et al.'s linearization and the more general PVST, which are thoroughly described in Chap. 10.

11.6.3 On Upwinding and Numerical Dispersion

Although the above matrix systems (11.37) for the discretized ADE's of species mass and heat transport are written by using the Galerkin method, they can easily be combined with upwind strategies which have been thoroughly described in Sect. 8.14. Useful upwind strategies refer to the SU and FU methods (Sect. 8.14.3), SC method (Sect. 8.14.4) and PGLS method (Sect. 8.14.5), in which the tensor of mechanical dispersion $\boldsymbol{D}_{\text{mech}}$ as part of the hydrodynamic dispersion tensor $\boldsymbol{D}_k$ and tensor of hydrodynamic thermodispersion $\boldsymbol{\Lambda}$ (11.2) is appropriately modified by stabilization terms in dependence on the actual spatial and temporal discretizations or solution gradients. Practically, the following schemes can be chosen if evaluating $\boldsymbol{D}^e_{\text{mech}}$ for each element e according to Table 11.3.

The temporal and spatial (upwind) discretization strategies affect the accuracy of the solution. The inherent truncation errors can be expressed in terms of numerical (nonphysical) dispersion D_{num} as described in Sect. 8.15 and summarized in Table 8.9. The original (physical) diffusivity is artificially raised by D_{num}. Typically, it is on element level

$$D_{\text{num}} = \alpha \frac{\|\boldsymbol{q}^e\|\, h^e}{2} + \Delta t_n {q^e}^2(\theta - \tfrac{1}{2}) \tag{11.43}$$

where $\alpha \in (1, 0)$ for upwind and GFEM (no upwind) scheme, respectively, and $\theta \in (\frac{1}{2}, 1)$ for TR and BE, respectively. As (11.43) shows, upwind schemes and fully implicit temporal approximations introduce a maximum numerical dispersion. Small element sizes h^e and time steps Δt_n are required to reduce D_{num}. Conse-

Table 11.3 Different upwind schemes by modified longitudinal and transverse dispersivities, β_L^e, β_T^e, respectively, occurring in the mechanical dispersion tensor $\boldsymbol{D}^e_{\text{mech}}$ of element e

	$\boldsymbol{D}^e_{\text{mech}} = \beta_T^e \lVert \boldsymbol{q}^e \rVert \boldsymbol{\delta} + (\beta_L^e - \beta_T^e) \frac{\boldsymbol{q}^e \otimes \boldsymbol{q}^e}{\lVert \boldsymbol{q}^e \rVert}$		
Scheme	β_L^e	β_T^e	Reference
GFEM (no upwind)[a]	β_L	β_T	
SU[b]	$\beta_L + \alpha \frac{h^e}{2}$	β_T	(8.245)
FU[b]	$\beta_L + \alpha \frac{h^e}{2}$	$\beta_T + \alpha \frac{h^e}{2}$	(8.251)
SC[c]	$\beta_L + \frac{1}{2} \alpha_c h^e \frac{\lVert \boldsymbol{q}^e_{\parallel} \rVert}{\lVert \boldsymbol{q}^e \rVert}$	$\beta_T + \frac{1}{2} \alpha_c h^e \frac{\lVert \boldsymbol{q}^e_{\parallel} \rVert}{\lVert \boldsymbol{q}^e \rVert}$	(8.258)
PGLS[d]	$\beta_L + Cr\, h^e$	β_T	(8.295)

[a] True (physical) dispersivities are β_L and β_T

[b] Upwind parameter $\alpha = 1$; factor $\frac{1}{2}$ for linear elements; h^e is characteristic element length defined by (8.239)

[c] α_c defined by (8.259); projected flux vector $\boldsymbol{q}^e_{\parallel}$ defined by (8.253)

[d] Courant number Cr defined in (8.236)

quently, a convection process is simulated with changed parameters, as quantified by an *effective* Rayleigh number

$$\text{Ra}^{\text{eff}} = \frac{\text{Ra}}{1 + \alpha Pg + (\theta - \frac{1}{2}) Cr\, Pg} \tag{11.44}$$

where Ra is the true (physical) Rayleigh number, Pg is the mesh Péclet number (8.236) and Cr is the Courant number (8.236). From (11.44) the danger from upwind schemes and/or fully implicit schemes becomes obvious: the solution can evolve to a point rather far from the real physics of a flow problem, if upwinding on coarse meshes, and/or fully implicit time marching schemes with large step sizes are used (cf. discussions in Sect. 8.14.1).

11.6.4 Preferred Strategy for Solving the Coupled Nonlinear Spatio-Temporarily Discretized System

In general, for the present class of transient nonlinear density-coupled flow and transport processes it cannot be predicted which time steps are allowable with respect to the accuracy requirements. Accordingly, a time marching recurrence scheme such as the $\theta-$method (Sect. 8.13.4) with predefined (fixed) time step sizes Δt_n is usually rather inappropriate and inefficient.[4] Our favorite method for transient variable-density problems is the GLS predictor-corrector time integrator (Sect. 8.13.5), which provides a cost-effective, robust and accurate technique in that

[4] The time integration of (11.38) by using the simple $\theta-$method (Sect. 8.13.4) gives

the time step size is increased whenever possible and decreased only when necessary due to the error estimates. In addition, the predictor-corrector method with the error-controlled adaptive time stepping is superior to linearize the nonlinear matrix equations for flow, mass and heat transport and allows the embedding of the one-step Newton (or alternatively one-step Picard) iteration method without the necessity of repeated iteration within each time step (cf. Sect. 8.18.4).

The GLS predictor-corrector solution strategy breaks down into the following main working steps:

STEP 0: Initialization

The predictor-corrector procedure necessitates the knowledge of the initial time derivative (history vector) $\dot{\boldsymbol{U}}_0$ of the state vector $\boldsymbol{U}(t)$ (11.40) containing the vectors of hydraulic head $\boldsymbol{h}$, species concentration $\boldsymbol{C}_k$ and temperature $\boldsymbol{T}$ at the nodal points. It can be solved by evaluating the matrix equation (11.38) under utilizing the IC's (11.6) as

$$\boldsymbol{G}(\boldsymbol{U}_0)\cdot\dot{\boldsymbol{U}}_0 = -\boldsymbol{K}(\boldsymbol{U}_0)\cdot\boldsymbol{U}_0 + \boldsymbol{Q}(\boldsymbol{U}_0) \tag{11.45}$$

where $\boldsymbol{U}_0^T = (\boldsymbol{h}_0 \; \boldsymbol{C}_{k,0} \; \boldsymbol{T}_0)$ is known at initial time. The system (11.43) needs to be solved only once and the extra work for the initialization is amortized over the rest of computation.

STEP 1: Predictor solutions

Perform explicit predictor solutions by using the 1st-order accurate FE and 2nd-order accurate AB scheme, respectively,

$$\begin{aligned}&\left(\frac{\boldsymbol{G}(\boldsymbol{U}_{n+1})}{\Delta t_n} + \boldsymbol{K}(\boldsymbol{U}_{n+1})\theta\right)\cdot\boldsymbol{U}_{n+1} = \\ &\left(\frac{\boldsymbol{G}(\boldsymbol{U}_{n+1})}{\Delta t_n} - \boldsymbol{K}(\boldsymbol{U}_{n+1})(1-\theta)\right)\cdot\boldsymbol{U}_n + \big(\boldsymbol{Q}(\boldsymbol{U}_{n+1})\theta + \boldsymbol{Q}(\boldsymbol{U}_n)(1-\theta)\big)\end{aligned}$$

where $\theta \in (\frac{1}{2}, 1)$ for the Crank-Nicolson and the fully implicit scheme,respectively. A nonlinear matrix system $\boldsymbol{R}_{n+1} = \boldsymbol{A}(\boldsymbol{U}_{n+1})\cdot\boldsymbol{U}_{n+1} - \boldsymbol{Z}(\boldsymbol{U}_{n+1}, \boldsymbol{U}_n) = \boldsymbol{0}$ results, which must be iteratively solved either via the Picard method (Sect. 8.18.1)

$$\boldsymbol{A}(\boldsymbol{U}_{n+1}^{\tau})\cdot\boldsymbol{U}_{n+1}^{\tau+1} = \boldsymbol{Z}(\boldsymbol{U}_{n+1}^{\tau}, \boldsymbol{U}_n) \quad \tau = 0, 1, 2, \ldots$$

or via the Newton method (Sect. 8.18.2)

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{U}_{n+1}^{\tau})\cdot\Delta\boldsymbol{U}_{n+1}^{\tau} &= -\boldsymbol{R}_{n+1}(\boldsymbol{U}_{n+1}^{\tau}, \boldsymbol{U}_n) \quad \tau = 0, 1, 2, \ldots\\ \Delta\boldsymbol{U}_{n+1}^{\tau} &= \boldsymbol{U}_{n+1}^{\tau+1} - \boldsymbol{U}_{n+1}^{\tau}\\ \boldsymbol{J}(\boldsymbol{U}_{n+1}^{\tau}) &= \frac{\partial\boldsymbol{R}_{n+1}(\boldsymbol{U}_{n+1}^{\tau}, \boldsymbol{U}_n)}{\partial\boldsymbol{U}_{n+1}^{\tau}}\end{aligned}$$

until satisfactory convergence is achieved for the iterations τ at each given time stage $n+1$. Note that this iterative solution strategy is also applicable to *steady-state* variable-density problems if setting $\theta = 1$ and $\Delta t_n \rightarrow \infty$.

$$\boldsymbol{U}_{n+1}^{p} = \begin{cases} \boldsymbol{U}_n + \Delta t_n \dot{\boldsymbol{U}}_n & \text{FE predictor} \\ \boldsymbol{U}_n + \frac{\Delta t_n}{2}\big[\big(2 + \frac{\Delta t_n}{\Delta t_{n-1}}\big)\dot{\boldsymbol{U}}_n - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\boldsymbol{U}}_{n-1}\big] & \text{AB predictor} \end{cases} \tag{11.46}$$

where the superposed p denotes the predictor value $\boldsymbol{U}_{n+1}^{p\,T}{=}(\boldsymbol{h}_{n+1}^{p}\;\boldsymbol{C}_{k,n+1}^{p}\;\boldsymbol{T}_{n+1}^{p})$ at the new time plane $n+1$. Note that, since $\dot{\boldsymbol{U}}_{n-1}$ is required, the AB formula cannot be applied before the second step ($n = 1$). The prediction has to be started with the FE scheme, where $\dot{\boldsymbol{U}}_0$ is available from (11.45).

STEP 2: Corrector solutions

Do corrector solution for the nonlinear matrix system (11.38) achieved by the TR or BE scheme

$$\Big(\frac{\boldsymbol{G}(\boldsymbol{U}_{n+1}^{p})}{\theta \Delta t_n} + \boldsymbol{K}(\boldsymbol{U}_{n+1}^{p}) + \hat{\boldsymbol{J}}(\boldsymbol{U}_{n+1}^{p})\Big) \cdot \boldsymbol{U}_{n+1} =$$

$$\boldsymbol{G}(\boldsymbol{U}_{n+1}^{p}) \cdot \Big[\frac{\boldsymbol{U}_n}{\theta \Delta t_n} + \big(\tfrac{1}{\theta} - 1\big)\dot{\boldsymbol{U}}_n\Big] + \hat{\boldsymbol{J}}(\boldsymbol{U}_{n+1}^{p}) \cdot \boldsymbol{U}_{n+1}^{p} + \boldsymbol{Q}(\boldsymbol{U}_{n+1}^{p}) \tag{11.47}$$

to determine $\boldsymbol{U}_{n+1}^{T} = (\boldsymbol{h}_{n+1}\;\boldsymbol{C}_{k,n+1}\;\boldsymbol{T}_{n+1})$ at the new time plane $n+1$, where $\theta \in (\frac{1}{2}, 1)$ for the TR and BE scheme, respectively. In (11.47) the predictor solution (11.46) is used to linearize the nonlinear dependencies of the matrix system and

$$\hat{\boldsymbol{J}}(\boldsymbol{U}_{n+1}^{p}) = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{J}}_k(\boldsymbol{h}_{n+1}^{p}, \boldsymbol{C}_{k,n+1}^{p}, \boldsymbol{T}_{n+1}^{p}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{J}}_T(\boldsymbol{h}_{n+1}^{p}, \boldsymbol{C}_{k,n+1}^{p}, \boldsymbol{T}_{n+1}^{p}) \end{pmatrix} \tag{11.48}$$

appears as the partial Jacobian matrix based on the predictor which results from the (one-step) Newton approach (note that this partial Jacobian does not exist for the Picard method). Basically, the matrix system (11.47) has to be solved simultaneously for $\boldsymbol{h}_{n+1}$, $\boldsymbol{C}_{k,n+1}$ and $\boldsymbol{T}_{n+1}$ if subjected to a complete iteration loop and to attain an improved rate of convergence for a full Newton approach, however, at the cost of a significant memory and computational burden, in particular for 3D and multispecies transport problems. The matrix system for such a simultaneous solution can be ill-conditioned due to the significantly different scales of the processes involved. But, in the preferred one-step predictor-linearized formulation the system (11.47) naturally decouples into the three partial systems

$$
\begin{aligned}
&\Big(\tfrac{\boldsymbol{O}(\boldsymbol{h}^p_{n+1})}{\theta\Delta t_n} + \tfrac{\boldsymbol{B}}{\theta\Delta t_n}\cdot\tfrac{\partial \boldsymbol{s}(\boldsymbol{h}^p_{n+1})}{\partial \boldsymbol{h}^p_{n+1}} + \boldsymbol{S}(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})\Big)\cdot\boldsymbol{h}_{n+1} = \\
&\qquad \Big(\boldsymbol{O}(\boldsymbol{h}^p_{n+1}) + \boldsymbol{B}\cdot\tfrac{\partial \boldsymbol{s}(\boldsymbol{h}^p_{n+1})}{\partial \boldsymbol{h}^p_{n+1}}\Big)\cdot\Big[\tfrac{\boldsymbol{h}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{h}}_n\Big] + \boldsymbol{F}(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1}) \\[2ex]
&\Big(\tfrac{\boldsymbol{H}_k(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1})}{\theta\Delta t_n} + \boldsymbol{E}_k(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1}) + \hat{\boldsymbol{J}}_k(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})\Big)\cdot\boldsymbol{C}_{k,n+1} = \\
&\qquad \boldsymbol{H}_k(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1})\cdot\Big[\tfrac{\boldsymbol{C}_{k,n}}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{C}}_{k\,n}\Big] + \hat{\boldsymbol{J}}_k(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})\cdot\boldsymbol{C}^p_{k,n+1} + \boldsymbol{R}_k \\[2ex]
&\Big(\tfrac{\boldsymbol{P}(\boldsymbol{h}^p_{n+1})}{\theta\Delta t_n} + \boldsymbol{L}(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1}) + \hat{\boldsymbol{J}}_T(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})\Big)\cdot\boldsymbol{T}_{n+1} = \\
&\qquad \boldsymbol{P}(\boldsymbol{h}^p_{n+1})\cdot\Big[\tfrac{\boldsymbol{T}_n}{\theta\Delta t_n} + \big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{T}}_n\Big] + \hat{\boldsymbol{J}}_T(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})\cdot\boldsymbol{T}^p_{n+1} + \boldsymbol{W}
\end{aligned}
\tag{11.49}
$$

for determining the corrector solutions $\boldsymbol{h}_{n+1}$, $\boldsymbol{C}_{k,n+1}$ and $\boldsymbol{T}_{n+1}$ in a *sequential iterative approach* (SIA), which reduces significantly the computational effort. The resulting linearized systems (11.49) consisting of algebraic equations of symmetric and unsymmetric structure are solved by techniques as described in Sect. 8.17. The partial Jacobians $\hat{\boldsymbol{J}}_k(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})$ and $\hat{\boldsymbol{J}}_T(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})$ appearing in (11.49) for the Newton method result from the nonlinearities of fluid density in the advective terms of matrices $\boldsymbol{E}_k$ and $\boldsymbol{L}$, respectively, and are given by[5]

$$
\begin{aligned}
\hat{\boldsymbol{J}}_k &= \frac{\partial \boldsymbol{E}_k(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})}{\partial \boldsymbol{C}^p_{k,n+1}}\cdot\boldsymbol{C}^p_{k,n+1} \\
\hat{J}_{k,ij} &= \sum_l \frac{\partial E_{k,il}}{\partial C^p_{k,j,n+1}} C^p_{k,l,n+1} \\
&= -\sum_e \int_{\Omega^e} \big(k^e_r \boldsymbol{K}^e f^e_\mu \cdot \boldsymbol{e}\big)\cdot N_i N_j \sum_k \beta^e_{c_k} \sum_l (\nabla N_l C^p_{k,l})\, d\Omega^e
\end{aligned}
\tag{11.50}
$$

and

$$
\begin{aligned}
\hat{\boldsymbol{J}}_T &= \frac{\partial \boldsymbol{L}(\boldsymbol{h}^p_{n+1},\boldsymbol{C}^p_{k,n+1},\boldsymbol{T}^p_{n+1})}{\partial \boldsymbol{T}^p_{n+1}}\cdot\boldsymbol{T}^p_{n+1} \\
\hat{J}_{T,ij} &= \sum_l \frac{\partial L_{il}}{\partial T^p_{j,n+1}} T^p_{l,n+1} \\
&= \sum_e \int_{\Omega^e} \big(k^e_r \boldsymbol{K}^e f^e_\mu \cdot \boldsymbol{e}\big)\cdot N_i N_j \beta^e(T^e) \sum_l (\nabla N_l T^p_l)\, d\Omega^e
\end{aligned}
\tag{11.51}
$$

[5] For the divergence form of the governing species mass and heat ADE's it results

$$
\begin{aligned}
\hat{J}_{k,ij} &= \sum_e \int_{\Omega^e} \big(k^e_r \boldsymbol{K}^e f^e_\mu \cdot \boldsymbol{e}\big)\cdot \nabla N_i N_j \sum_k \beta^e_{c_k} \sum_l (N_l C^p_{k,l})\, d\Omega^e \\
\hat{J}_{T,ij} &= -\sum_e \int_{\Omega^e} \big(k^e_r \boldsymbol{K}^e f^e_\mu \cdot \boldsymbol{e}\big)\cdot \nabla N_i N_j \beta^e(T^e) \sum_l (N_l T^p_l)\, d\Omega^e
\end{aligned}
$$

where for convenience other nonlinear dependencies, e.g., occurring in the velocity-dependent dispersion, fluid viscosity or saturation, are not incorporated. Note that a Newton scheme for the flow matrix system under unsaturated conditions can be alternatively used as described in Sects. 10.6.2 and 10.7.

STEP 3: Updated accelerations

Update the new acceleration vectors by inverting the FE and BE, respectively, according to Table 8.7 as

$$\dot{\boldsymbol{U}}_{n+1} = \begin{cases} \frac{\boldsymbol{U}_{n+1}-\boldsymbol{U}_n}{\Delta t_n} & \text{FE} \\ \left(2 - \frac{\Delta t_{n-1}}{\Delta t_n + \Delta t_{n-1}}\right)\left(\frac{\boldsymbol{U}_{n+1}-\boldsymbol{U}_n}{\Delta t_n}\right) - \left(\frac{\Delta t_n}{\Delta t_n + \Delta t_{n-1}}\right)\left(\frac{\boldsymbol{U}_n-\boldsymbol{U}_{n-1}}{\Delta t_{n-1}}\right) & \text{AB} \end{cases} \tag{11.52}$$

to obtain $\dot{\boldsymbol{U}}_{n+1}^T = (\dot{\boldsymbol{h}}_{n+1}\ \dot{\boldsymbol{C}}_{k,n+1}\ \dot{\boldsymbol{T}}_{n+1})$ at the new time plane $n+1$.

STEP 4: Error estimation

Compute the LTE for the FE/BE and AB/TR scheme as a function of the corrector and predictor solutions in the form (cf. Table 8.7)

$$\begin{aligned} \boldsymbol{d}_{n+1}^h &= \varphi(\boldsymbol{h}_{n+1} - \boldsymbol{h}_{n+1}^p) \\ \boldsymbol{d}_{n+1}^{ck} &= \varphi(\boldsymbol{C}_{k,n+1} - \boldsymbol{C}_{k,n+1}^p) \quad (k = 1, \dots, N) \\ \boldsymbol{d}_{n+1}^t &= \varphi(\boldsymbol{T}_{n+1} - \boldsymbol{T}_{n+1}^p) \end{aligned} \tag{11.53}$$

with

$$\varphi = \begin{cases} \frac{1}{2} & \text{for FE/BE} \\ \frac{1}{3\left(1+\frac{\Delta t_{n-1}}{\Delta t_n}\right)} & \text{for AB/TR} \end{cases} \tag{11.54}$$

Appropriate error norms are applied to the LTE vectors $\boldsymbol{d}_{n+1}^h$, $\boldsymbol{d}_{n+1}^{ck}$ and $\boldsymbol{d}_{n+1}^t$. Commonly, the weighted RMS L_2 error norms

$$\begin{aligned} \|\boldsymbol{d}_{n+1}^h\|_{L_2} &= \left[\frac{1}{N_P}\left(\sum_{i=1}^{N_P}\left|\frac{d_{i,n+1}^h}{h_{\max,n+1}}\right|^2\right)\right]^{1/2} \\ \|\boldsymbol{d}_{n+1}^{ck}\|_{L_2} &= \left[\frac{1}{N_P}\left(\sum_{i=1}^{N_P}\left|\frac{d_{i,n+1}^{ck}}{C_{k,\max,n+1}}\right|^2\right)\right]^{1/2} \quad (k = 1, \dots, N) \\ \|\boldsymbol{d}_{n+1}^t\|_{L_2} &= \left[\frac{1}{N_P}\left(\sum_{i=1}^{N_P}\left|\frac{d_{i,n+1}^t}{T_{\max,n+1}}\right|^2\right)\right]^{1/2} \end{aligned} \tag{11.55}$$

and the maximum L_∞ error norms

$$\begin{aligned}\|\boldsymbol{d}^h_{n+1}\|_{L_\infty} &= \tfrac{1}{h_{\max,n+1}}\max_i|d^h_{i,n+1}| \\ \|\boldsymbol{d}^{ck}_{n+1}\|_{L_\infty} &= \tfrac{1}{C_{k,\max,n+1}}\max_i|d^{ck}_{i,n+1}| \quad (k=1,\dots,N) \\ \|\boldsymbol{d}^t_{n+1}\|_{L_\infty} &= \tfrac{1}{T_{\max,n+1}}\max_i|d^t_{i,n+1}|\end{aligned} \tag{11.56}$$

are chosen, where $h_{\max,n+1}$, $C_{k,\max,n+1}$ and $T_{\max,n+1}$ correspond to the maximum values of hydraulic head, species concentration and temperature, respectively, detected at the time plane $n+1$, and used to normalize the solution vectors.

STEP 5: Tactic of time stepping

The potential size of the next time step can be computed by means of the error estimates (11.53), (11.55), (11.56), the current time step size Δt_n, and a user-specified error tolerance ϵ as

$$\Delta t_{n+1} = \Delta t_n \left(\frac{\epsilon}{\max(\|\boldsymbol{d}^h_{n+1}\|_{L_p}, \|\boldsymbol{d}^{c1}_{n+1}\|_{L_p}, \dots, \|\boldsymbol{d}^{cN}_{n+1}\|_{L_p}, \|\boldsymbol{d}^t_{n+1}\|_{L_p})}\right)^{1/\lambda} \tag{11.57}$$

where

$$\begin{aligned}\lambda &= \begin{cases} 2 & \text{for FE/BE} \\ 3 & \text{for AB/TR}\end{cases} \\ p &= \begin{cases} 2 & \text{for RMS error norm} \\ \infty & \text{for maximum error norm}\end{cases}\end{aligned} \tag{11.58}$$

and $\boldsymbol{d}^h_{n+1}$, $\boldsymbol{d}^{ck}_{n+1}$ $(k=1,\dots,N)$ and $\boldsymbol{d}^t_{n+1}$ are defined by (11.53). To monitor the progress of the solution we use the criteria as summarized in Table 8.7.

11.7 Consistent Velocity Approximation

In the Darcy law (11.1) the discretization of the fluxes (velocities) $\boldsymbol{q}$ is nontrivial if density effects become important. Specifically, a lower-order approximation attainable for the hydraulic head gradients ∇h can conflict with a high-order spatial variation in the gravity (buoyancy) term $\chi \boldsymbol{e}$ due to the following reasons. There may be significant parts of the domain where the Darcy velocity should be zero (or very small). In these regions, there should be a balance between two contributions to the Darcy velocity: the hydraulic head gradient term and the density term. However, if the hydraulic head, the species concentration and the temperature are approximated using finite-element approximations based on the same order of polynomials, then the hydraulic head gradient term is a lower-order polynomial in position, and therefore cannot in general match the variation with position of the gravity term, which varies with position in the same way as the species concentration and temperature. As a result, although the Darcy velocity may be zero in some average sense over an element, it varies on the scale of the elements. These artificial

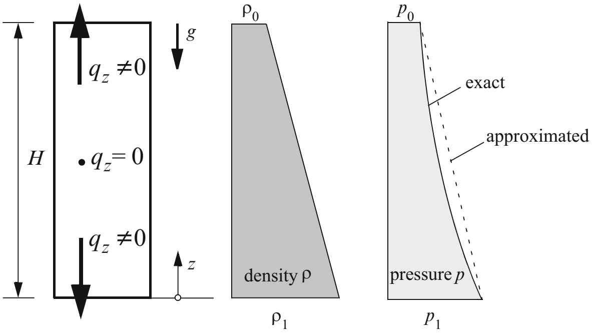


Fig. 11.10 Hydrostatic condition in a finite element of height H under a linear density gradient $\rho = \rho_1 - (\rho_1 - \rho_0)\frac{z}{H}$; spurious vertical velocities q_z caused by an inexact pressure (or hydraulic head) approximation

variations may have a major effect on the computed mass and heat transport, both because they may lead to spurious advection of mass and heat and also because they may lead to an increased dispersivity.

This problem has been addressed by Voss [550], Voss and Souza [552], Herbert et al. [244] and Leijnse [336] who proposed modified schemes, termed *consistent velocity approximation*, for evaluation of the discontinuous derivatives. In Voss and Souza's approach the spatial variation in the gravity term is reduced to the same spatial variation as occurring in the hydraulic head gradient, i.e., for linear finite elements the hydraulic head gradient is constant (piecewise constant per element) and accordingly, the gravity term should also be piecewise constant. While Voss and Souza [552] and Leijnse [336] tried to overcome the problem of consistency by precision reduction, Herbert et al. [244] solved it by introducing a mixed higher-order approximation for the different Darcy flux terms, which significantly raised the computational expense.

11.7.1 The Hydrostatic Condition: The Requirement of Consistency

Consider a hydrostatic situation for a single finite element as shown in Fig. 11.10, where the fluid density ρ varies linearly in the vertical z−direction:

$$\rho = \rho_1 - (\rho_1 - \rho_0)\tfrac{z}{H}, \qquad 0 \leq z \leq H \tag{11.59}$$

Under such a hydrostatic condition we require that the Darcy velocity $\boldsymbol{q}$ expressed in its h−formulation (11.1) as $\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})$ or in its p−formulation (3.258) as $\boldsymbol{q} = -\frac{k_r \boldsymbol{K}}{\mu} \cdot (\nabla p - \rho \boldsymbol{g})$ must be zero everywhere. This is termed as the *requirement of consistency*, which implies

$$\boldsymbol{q} \equiv \boldsymbol{0}, \quad \nabla h = -\chi \boldsymbol{e}, \quad \nabla p = \rho \boldsymbol{g} \tag{11.60}$$

Suppose that the gravity is directed vertically along the z−coordinate so that $\boldsymbol{g}^T = (0\ 0\ -g)$ and $\boldsymbol{e}^T = (0\ 0\ 1)$ the pressure p and hydraulic head h, respectively,

$$p(z) = -g \int_{z_1}^{z} \rho(z) dz, \quad h(z) = -\int_{z_1}^{z} \chi(z) dz, \quad \chi(z) = \tfrac{\rho(z)-\rho_0}{\rho_0} \tag{11.61}$$

yield for the *linear* density function (11.59) a *quadratic* shape of the pressure as

$$p = p_1 - g\big(\rho_1 z - \tfrac{\rho_1-\rho_0}{2H} z^2\big) \tag{11.62}$$

and similarly for the hydraulic head as

$$h = h_1 - \tfrac{\rho_1-\rho_0}{\rho_0}\big(z - \tfrac{z^2}{2H}\big) \tag{11.63}$$

11.7.2 The Artifact: Spurious Nonconsistent Velocities and Common Ways to Overcome

Typically, in a discretization algorithm the species concentration C_k and/or the temperature T is linearly approximated in a finite element. This leads to a corresponding linear relationship for the density ρ or buoyancy coefficient χ as considered above. But, the hydraulic head is also approximated by a linear function in an element. This is (in the example of Fig. 11.10):

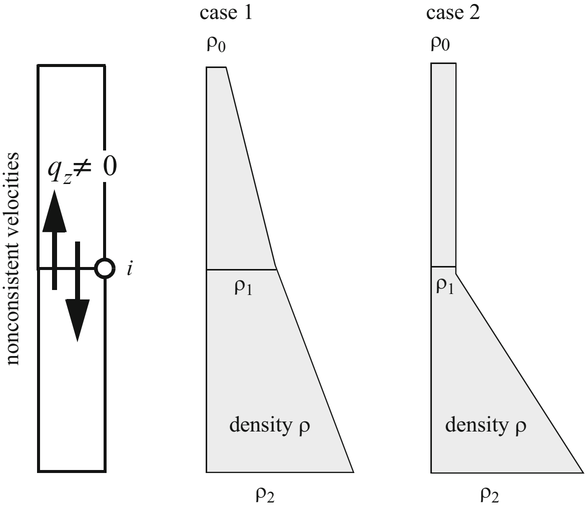
$$h \approx \hat{h} = h_0 - (h_1 - h_0)\tfrac{z}{H}, \qquad 0 \leq z \leq H \tag{11.64}$$

Inserting (11.64) into the Darcy equation (11.1) and using the exact nodal values h_1 and $h_0 = h_1 - \frac{\rho_1-\rho_0}{\rho_0}\frac{H}{2}$ from (11.63) as well as for the density (11.59) we get for the z−component of the approximated velocity

$$q_z = -k_r K_{zz} f_\mu \big[\tfrac{\rho_1-\rho_0}{\rho_0}(\tfrac{1}{2} - \tfrac{z}{H})\big], \qquad 0 \leq z \leq H \tag{11.65}$$

It clearly indicates that the approximated velocity only vanishes at the middle position ($z = H/2$) while at the other points *artificial nonzero quantities* occur which take maximum values with opposite signs at the left ($z = 0$) and right ($z = H$) point (cf. Fig. 11.10). Those *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 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 mechanical dispersion tensor $\boldsymbol{D}_{\text{mech}}$ (11.2) at element level an artificial increase of hydrodynamic dispersion (mixing) can result [336].

Fig. 11.11 Continuous nodal velocity by averaging (smoothing) nonconsistent velocities for two cases of vertical density profiles



The most important way to overcome the problem is in reducing the spatial variability in the gravity (buoyancy) term. Commonly, the gravity term is averaged in the appropriate direction so as proposed by Voss [550], Voss and Souza [552] and Leijnse [336]. In the above example we have to use now $\rho = (\rho_1 + \rho_0)/2$ and find with the exact nodal values h_1 and $h_0 = h_1 - \frac{\rho_1 - \rho_0}{\rho_0}\frac{H}{2}$:

$$q_z = -k_r K_{zz} f_\mu \big(\underbrace{-\tfrac{\rho_1 - \rho_0}{2\rho_0}}_{\nabla_z h} + \underbrace{\tfrac{\rho_1 - \rho_0}{2\rho_0}}_{\chi e_z} \big) = 0 \tag{11.66}$$

which satisfies the equilibrium at all points.

Another possibility is in averaging the nonconsistent velocities at nodal points by the local or global smoothing techniques as thoroughly described in Sect. 8.19. It may smooth out the spurious velocities. Let us consider the following situations as shown in Fig. 11.11, where node i shares two finite elements.

The smoothing procedure for the nonconsistent velocity (11.65) leads to a velocity at the node i as

$$q_{z,i} = k_r K_{zz} f_\mu \frac{1}{2\rho_0}\left(\frac{\rho_0 + \rho_2}{2} - \rho_1 \right) \tag{11.67}$$

If we can assume that the density ρ_1 at the node i is an average of the upper and lower density values, i.e., $\rho_1 = \frac{\rho_0 + \rho_2}{2}$, then the nodal velocity (11.67) becomes consistent with $q_{z,i} = 0$. Obviously, this is true (or approximately true) for typical density profiles as shown as case 1 in Fig. 11.11. 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. 11.11 at the node i where an upgoing spurious velocity remains in order of

$$q_{z,i} = k_r K_{zz} f_\mu \frac{1}{2\rho_0}\left(\frac{\rho_2 - \rho_1}{2}\right) \tag{11.68}$$

and the consistency is not satisfied at the node under those 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 (11.60) 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. Hence, smoothing is a procedure to derive continuous nodal velocities which are often, but not always consistent in the sense of the statement (11.60).
4. In the context of variable-density flow a more general procedure is required for consistent velocity approximations which will be described next.

11.7.3 The Frolkovič-Knabner Algorithm

Frolkovič [177] and Knabner and Frolkovič [312] introduced an algorithm, hereafter referred to as the *Frolkovič-Knabner algorithm* (FKA), to approximate consistent velocities in 2D and 3D finite elements in a more general manner. FKA is described for isoparametric families of finite elements, where the computations are realized on generalized (local) coordinates $\boldsymbol{\eta}$ (8.68), cf. Sect. 8.8.1. The idea is the introduction of integral functions $H^e_\xi(\boldsymbol{\eta}), H^e_\eta(\boldsymbol{\eta}), H^e_\zeta(\boldsymbol{\eta})$ to evaluate the Darcy velocities $\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})$ for each element e in the form:

$$\boldsymbol{q}(\boldsymbol{x}(\boldsymbol{\eta})) \to \boldsymbol{q}^e(\boldsymbol{\eta}_p) = -\sum_J^{N_{\mathrm{BN}}} k_r^e \boldsymbol{K}^e f_\mu^e \cdot \left((\boldsymbol{J}^e)^{-1} \cdot \begin{pmatrix} (h^e_J + H^e_{\xi J}) \frac{\partial}{\partial \xi} N^e_J(\boldsymbol{\eta}_p) \\ (h^e_J + H^e_{\eta J}) \frac{\partial}{\partial \eta} N^e_J(\boldsymbol{\eta}_p) \\ (h^e_J + H^e_{\zeta J}) \frac{\partial}{\partial \zeta} N^e_J(\boldsymbol{\eta}_p) \end{pmatrix} \right) \quad (p = 1, \ldots, m) \tag{11.69}$$

where $\boldsymbol{\eta}_p$ is the vector of local coordinates, e.g., (ξ_p, η_p, ζ_p) for a 3D element, $(\boldsymbol{J}^e)^{-1}$ is the inverse Jacobian (8.119) of the isoparametric element, $H^e_{\xi J}, H^e_{\eta J}, H^e_{\zeta J}$ are the nodal integral functions at local node J and m is the total number of Gauss points, cf. Sects. 8.8.2 and 8.19.2. The integral functions are derived in Appendix K.

By using these integral functions $H^e_\xi(\boldsymbol{\eta})$, $H^e_\eta(\boldsymbol{\eta})$, $H^e_\zeta(\boldsymbol{\eta})$ the same spatial variability for both the hydraulic head gradient term and the buoyancy term are achieved to ensure consistent velocities.

The element-by-element evaluation (11.69) performed at the Gauss points p for each element e leads naturally to a consistent velocity field, which is in general discontinuous at the nodal points. To obtain continuous velocities local or global smoothing techniques as thoroughly described in Sect. 8.19 can be easily applied. Obviously, the smoothing procedures have no effect on the consistency of the velocity. Since the velocities $\boldsymbol{q}^e(\boldsymbol{\eta}_{p\to J})$ for each element e are always consistent at a node J, a smoothed (continuous) velocity must be consistent too.

11.8 Flow, Species Mass and Heat Budget Evaluation

To obtain precise budget evaluations for flow, species mass and heat the CBFM, as introduced in Sect. 8.19.2, is applied to the specific weak formulations of the coupled equation system. The corresponding boundary fluxes on Γ have to be evaluated from the basic weak statements (11.32)–(11.34) written as

$$\begin{aligned}\int_\Gamma N_i\, q_{n_h}\, d\Gamma = -\int_\Omega N_i\, sS_o \frac{\partial h}{\partial t} d\Omega - \int_\Omega N_i \varepsilon \frac{\partial s}{\partial t} d\Omega -\\ \int_\Omega \nabla N_i \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega +\\ \int_\Omega N_i (Q_h + Q_{\mathrm{EOB}}) d\Omega - Q_w(t)|_i\end{aligned} \tag{11.70}$$

$$\begin{aligned}\int_\Gamma N_i\, q_{n_{kC}}\, d\Gamma = -\int_\Omega N_i \varepsilon s \acute{\Re}_k \frac{\partial C_k}{\partial t} d\Omega - \int_\Omega N_i \boldsymbol{q} \cdot \nabla C_k d\Omega -\\ \int_\Omega \nabla N_i \cdot (\boldsymbol{D}_k \cdot \nabla C_k) d\Omega - \int_\Omega N_i [(\varepsilon s \vartheta_k \Re_k + Q_h) C_k - \hat{R}_k - Q_k] d\Omega -\\ (C_{kw} - C_k) Q_w(t)|_i\end{aligned} \tag{11.71}$$

$$\begin{aligned}\int_\Gamma N_i\, q_{n_T}\, d\Gamma = -\int_\Omega N_i \big(\varepsilon s \rho c + (1-\varepsilon)\rho^s c^s\big) \frac{\partial T}{\partial t} d\Omega -\\ \int_\Omega N_i \rho c \boldsymbol{q} \cdot \nabla T d\Omega - \int_\Omega \nabla N_i \cdot (\boldsymbol{\Lambda} \cdot \nabla T) d\Omega -\\ \int_\Omega N_i [\rho c Q_h (T - T_0) - Q_T] d\Omega - \rho c (T_w - T) Q_w(t)|_i\end{aligned} \tag{11.72}$$

where all BC-related boundary segments are joined on $\varGamma$, q_{n_h}, $q_{n_{kC}}$ and q_{n_T} are the boundary fluxes for flow, species mass and heat, respectively. Expanding the boundary fluxes on $\varGamma$ as described in Sect. 8.19.2 the following matrix system results to solve the consistent boundary flux vector $\boldsymbol{q}_n$, viz.,

$$\boldsymbol{M}^{\dagger} \cdot \boldsymbol{q}_n = -\boldsymbol{G}(\boldsymbol{U}) \cdot \dot{\boldsymbol{U}} - \boldsymbol{K}^{\dagger}(\boldsymbol{U}) \cdot \boldsymbol{U} + \boldsymbol{Q}^{\dagger}(\boldsymbol{U}) \tag{11.73}$$

for known $\boldsymbol{U}$ and $\dot{\boldsymbol{U}}$ at the corresponding evaluation time t_{n+1}, where $\boldsymbol{U}$, $\dot{\boldsymbol{U}}$ and $\boldsymbol{G}(\boldsymbol{U})$ are defined in (11.39)–(11.42) and

$$\boldsymbol{q}_n = \begin{pmatrix} \boldsymbol{q}_{n_h} \\ \boldsymbol{q}_{n_{kC}} \\ \boldsymbol{q}_{n_T} \end{pmatrix}, \quad \boldsymbol{M}^{\dagger} = \begin{pmatrix} \boldsymbol{M} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{M} \end{pmatrix} \tag{11.74}$$

$$\begin{aligned} \boldsymbol{K}^{\dagger}(\boldsymbol{U}) &= \begin{pmatrix} \boldsymbol{S}^{\dagger}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E}_k^{\dagger}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{L}^{\dagger}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) \end{pmatrix} \\ \boldsymbol{Q}^{\dagger}(\boldsymbol{U}) &= \begin{pmatrix} \boldsymbol{F}^{\dagger}(\boldsymbol{h}, \boldsymbol{C}_k, \boldsymbol{T}) \\ \boldsymbol{R}_k^{\dagger} \\ \boldsymbol{W}^{\dagger} \end{pmatrix} \end{aligned} \tag{11.75}$$

with

$$\begin{aligned}
\boldsymbol{M} &= M_{ij} &&= \int_{\varGamma} N_i N_j \, d\varGamma \\
\boldsymbol{S}^{\dagger} &= S_{ij}^{\dagger} &&= \int_{\varOmega} \nabla N_i \cdot \big(k_r(s(h)) \boldsymbol{K} f_{\mu}(C_k, T) \cdot \nabla N_j\big) d\varOmega \\
\boldsymbol{E}_k^{\dagger} &= E_{ij,k}^{\dagger} &&= \int_{\varOmega} N_i \boldsymbol{q}(h, C_k, T) \cdot \nabla N_j \, d\varOmega + \int_{\varOmega} \nabla N_i \cdot (\boldsymbol{D}_k \cdot \nabla N_j) d\varOmega + \\
& && \quad \int_{\varOmega} (\varepsilon s(h) \vartheta_k \Re_k + Q_h) N_i N_j \, d\varOmega - \delta_{ij} Q_w(t)\big|_i \\
\boldsymbol{L}^{\dagger} &= L_{ij}^{\dagger} &&= \int_{\varOmega} N_i \rho c \boldsymbol{q}(h, C_k, T) \cdot \nabla N_j \, d\varOmega + \int_{\varOmega} \nabla N_i \cdot (\boldsymbol{\varLambda} \cdot \nabla N_j) d\varOmega + \\
& && \quad \int_{\varOmega} \rho c Q_h N_i N_j \, d\varOmega - \delta_{ij} \rho c Q_w(t)\big|_i \\
\boldsymbol{F}^{\dagger} &= F_i^{\dagger} &&= \int_{\varOmega} N_i \big(Q_h + Q_{\text{EOB}}(h, C_k, T)\big) d\varOmega - \\
& && \quad \int_{\varOmega} \nabla N_i \cdot \big(k_r(s(h)) \boldsymbol{K} f_{\mu}(C_k, T) \cdot \chi(C_k, T) \boldsymbol{e}\big) d\varOmega - Q_w(t)\big|_i \\
\boldsymbol{R}_k^{\dagger} &= R_{i,k}^{\dagger} &&= \int_{\varOmega} N_i (\hat{R}_k + Q_k) d\varOmega - C_{kw} Q_w(t)\big|_i \\
\boldsymbol{W}^{\dagger} &= W_i^{\dagger} &&= \int_{\varOmega} N_i Q_T \, d\varOmega + \int_{\varOmega} N_i \rho c T_0 Q_h \, d\varOmega - \rho c T_w Q_w(t)\big|_i
\end{aligned} \tag{11.76}$$

in which $(i, j = 1, \ldots, N_{\text{P}})$ and $(k = 1, \ldots, N)$. In the budget analysis the integral boundary balance flux is directly evaluated at each boundary node by

$$\boldsymbol{Q}_n = -\boldsymbol{M}^{\dagger} \cdot \boldsymbol{q}_n = \boldsymbol{G}(\boldsymbol{U}) \cdot \dot{\boldsymbol{U}} + \boldsymbol{K}^{\dagger}(\boldsymbol{U}) \cdot \boldsymbol{U} - \boldsymbol{Q}^{\dagger}(\boldsymbol{U}) \tag{11.77}$$

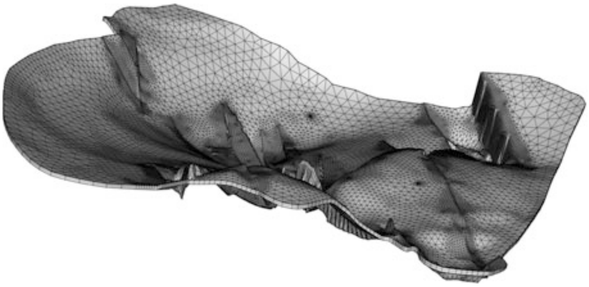


Fig. 11.12 Shape of a discretized thin aquifer (vertical exaggeration 5:1)

with

$$\boldsymbol{Q}_n = \begin{pmatrix} \boldsymbol{Q}_{n_h} \\ \boldsymbol{Q}_{n_{kC}} \\ \boldsymbol{Q}_{n_T} \end{pmatrix} \tag{11.78}$$

where $\boldsymbol{Q}_{n_h}$, $\boldsymbol{Q}_{n_{kC}}$ and $\boldsymbol{Q}_{n_T}$ correspond to the nodal vectors of the integral boundary fluxes for flow, species mass and heat, respectively.

Note that the boundary mass flux $q_{n_{kC}}$ and boundary heat flux q_{n_T} comprise only its dispersive and conductive part, respectively, in the convective forms of the governing transport equations. To obtain in addition the advective part of their boundary fluxes an auxiliary weak formulation must be applied such as described in Sect. 8.19.2.4. The expressions for the total (dispersive/conductive plus advective) boundary fluxes will be given in Sect. 12.4 for mass transport and in Sect. 13.4 for heat transport.

11.9 Modeling 2D Horizontally Schematized Aquifers Using Projected Gravity

11.9.1 2D Treatment of Thin, Slightly Sloped or Curved Aquifers

The numerical effort in solving variable-density flow is generally high due to the potential need for a suitably refined spatial and temporal discretization. This has serious consequences particularly in modeling of 3D problems, where meshes must be appropriately refined in all coordinate directions. In a 3D model, even if the aquifer is thin relative to its horizontal extent, a sufficient vertical discretization is usually required (Fig. 11.12).

However, there is a special case for which a fully 3D meshing of the problem can be avoided if the following conditions hold:

- There is a thin aquifer with an essentially horizontal (aquifer-type) flow for which the vertical flow components can be neglected. The horizontal extent of the aquifer is much larger compared to the aquifer thickness. Hence, flow and transport equations can be vertically integrated. This procedure is associated with the well-known Dupuit assumption [33] (cf. Sect. 3.5).
- The aquifer is slightly sloped or curved so that gravity can effect the movement of a solute (or heat) in such an aquifer.
- The aquifer is confined and saturated.

A typical application refers to the brine movement in a large-scale deep aquifer of a basin form. The brine moves down in deeper locations of the basin by gravity effects. The process is density-driven due to the sloped geometry of the aquifer layer. Under such conditions there is a way to model the variable-density solute distribution only in 2D. It is based on a 2D horizontally schematized aquifer described by vertically integrated equations and a projected gravity field.

11.9.2 Vertically Averaged Equations in a Confined Aquifer Including Gravity Term

In application of the averaging procedures described in Sect. 3.10.7 the following 2D vertically averaged flow, species mass and heat transport equations valid for a confined aquifer can be derived from (11.1) to (11.3)

$$\begin{aligned}
\bar{S}_o \frac{\partial h}{\partial t} + \nabla \cdot \bar{\boldsymbol{q}} &= \bar{Q} + \bar{Q}_{\text{EOB}} \\
\bar{\boldsymbol{q}} &= -\boldsymbol{T} f_\mu \cdot \big(\nabla h + \chi \boldsymbol{e}\big) \\
\varepsilon \bar{\acute{\mathfrak{R}}}_k \frac{\partial C_k}{\partial t} + \bar{\boldsymbol{q}} \cdot \nabla C_k - \nabla \cdot (\bar{\boldsymbol{D}}_k \cdot \nabla C_k) + (\varepsilon \vartheta_k \bar{\mathfrak{R}}_k + \bar{Q}) C_k &= \bar{\tilde{R}}_k \quad (k = 1, \ldots, N) \\
B\big(\varepsilon \rho c + (1-\varepsilon)\rho^s c^s\big) \frac{\partial T}{\partial t} + \rho c \bar{\boldsymbol{q}} \cdot \nabla T - \nabla \cdot (\bar{\boldsymbol{\Lambda}} \cdot \nabla T) &= \bar{H}_e - \rho c (T - T_0) \bar{Q}
\end{aligned} \tag{11.79}$$

associated with the constitutive relations

$$\begin{aligned}
\chi = \tfrac{\rho - \rho_0}{\rho_0} &= \sum_k \beta_{c_k} (C_k - C_{k0}) - \beta(T)(T - T_0) \\
\beta_{c_k} &= \tfrac{\alpha_k}{C_{ks} - C_{k0}} \\
\bar{\boldsymbol{D}}_k &= \varepsilon B D_k \boldsymbol{\delta} + \bar{\boldsymbol{D}}_{\text{mech}} \\
\bar{\boldsymbol{\Lambda}} &= B\big[\varepsilon \Lambda + (1-\varepsilon)\Lambda^s\big] \boldsymbol{\delta} + \rho c \bar{\boldsymbol{D}}_{\text{mech}} \\
\bar{\boldsymbol{D}}_{\text{mech}} &= \beta_T \|\bar{\boldsymbol{q}}\| \boldsymbol{\delta} + (\beta_L - \beta_T) \tfrac{\bar{\boldsymbol{q}} \otimes \bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|}
\end{aligned} \tag{11.80}$$

and the extended Oberbeck-Boussinesq term

$$\bar{Q}_{\text{EOB}} = -\bar{\boldsymbol{q}} \cdot \Big(\tfrac{\bar{S}_o}{\varepsilon B} \nabla h + \textstyle\sum_k \beta_{c_k} \nabla C_k - \beta^* \nabla T \Big) - \varepsilon B \Big(\textstyle\sum_k \beta_{c_k} \tfrac{\partial C_k}{\partial t} - \beta^* \tfrac{\partial T}{\partial t} \Big) \tag{11.81}$$

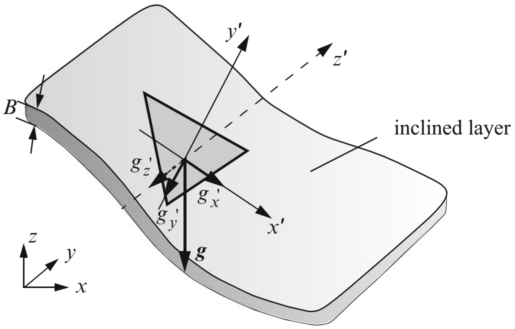


Fig. 11.13 Global $x - y - z$–coordinate system and local (rotated) $x' - y' - z'$–coordinate system for a finite element located on an inclined aquifer layer. Global gravity vector $\boldsymbol{g}$ is dissected by its local components $g_{x'}, g_{y'}, g_{z'}$

written for the convective form of transport equations, where $B = B(x, y)$ is the aquifer thickness, $\boldsymbol{T} = B\boldsymbol{K}$ is the tensor of transmissivity (3.302), f_μ, $\beta(T)$ and β^* are given in (11.2) and (11.3), respectively. The important difference to the standard formulation for 2D horizontal problems in confined aquifers (summarized in Table 3.11) is the buoyancy (gravity) term $\chi\boldsymbol{e}$ still appearing in the Darcy equation of (11.79). This term normally vanishes for a 'perfect' horizontal aquifer geometry because the gravity acts always perpendicular (vertical) to the aquifer horizon. However, if the aquifer is sloped or curved there are components of the gravity directed along the layer of the aquifer.

11.9.3 Local (Layer-Oriented) Coordinates $\boldsymbol{x}'$ and the Projected Gravity Term

Let us consider the situation of an inclined aquifer layer as shown in Fig. 11.13. We introduce local coordinates $\boldsymbol{x}'$ at a local point on the inclined aquifer in such a manner that x' and y' form the principal axes correlated with the geologic layer structure, while z' is directed perpendicular to the actual 2D $x' - y'$–computational plane.

The coordinate transformation between the global coordinates $\boldsymbol{x}$ and the local (layer-oriented) coordinates $\boldsymbol{x}'$ is described by the rotation matrix $\boldsymbol{A}$ as

$$\begin{gathered} \boldsymbol{x}' = \boldsymbol{A} \cdot \boldsymbol{x} \\ \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} \end{gathered} \tag{11.82}$$

Accordingly, the gravity components in the local coordinates are given by the transformation

$$
\begin{gathered}
\boldsymbol{g}' = \boldsymbol{A} \cdot \boldsymbol{g} \\
\begin{pmatrix} g_{x'} \\ g_{y'} \\ g_{z'} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \cdot \begin{pmatrix} g_x \\ g_y \\ g_z \end{pmatrix}
\end{gathered}
\tag{11.83}
$$

The rotation matrix $\boldsymbol{A}$ is performed for each finite element e as $\boldsymbol{A}^e$, which is described in Sect. 7.3.2. The components of $\boldsymbol{A}^e$ have the form:

$$
A^e_{ij} = \cos(\boldsymbol{u}_i, \boldsymbol{e}_j) = \frac{\boldsymbol{u}_i \cdot \boldsymbol{e}_j}{\|\boldsymbol{u}_i\| \|\boldsymbol{e}_j\|} \quad (i = 1, 2) \;\; (j = 1, 2, 3) \tag{11.84}
$$

with the base vectors (2.5) in 3D

$$
\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{11.85}
$$

where $\boldsymbol{u}_i$, (7.15)–(7.17), are directional vectors, which are evaluated for each finite element e in the 3D space.

The complete set of governing equations (11.79)–(11.81) is formulated in the local coordinates $\boldsymbol{x}'$, where the gravitational unit vector $\boldsymbol{e}'$ is also written for the local $x' - y'$–components directed along the principal axes of the inclined layer. They can be computed by the projection

$$
\begin{pmatrix} e_{x'} \\ e_{y'} \end{pmatrix} = \boldsymbol{e}' = -\frac{\boldsymbol{g}'}{\|\boldsymbol{g}'\|} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{11.86}
$$

Note, in (11.86) it is $\|\boldsymbol{g}'\| = \|\boldsymbol{g}\|$ and it is assumed that the gravity acts strictly downwards parallel to the global z–axis, i.e., $\boldsymbol{g}^T = (0\;0\;-g)$ and $\|\boldsymbol{g}\| = g$, where g is the gravitational acceleration constant.

In using this transformation procedure the 3D problem is mapped onto a 2D geometry so as exemplified for an idealized hemispherical basin geometry shown in Fig. 11.14. The variable-density effect is illustrated in Fig. 11.15 for this example. It shows how a dense solute sinks down to the center of the hemispherical basin in time caused by an exclusive action of gravity (i.e., accomplished by free convection). The density-driven solute movement is strongly dependent on the parameter heterogeneity so as indicated.

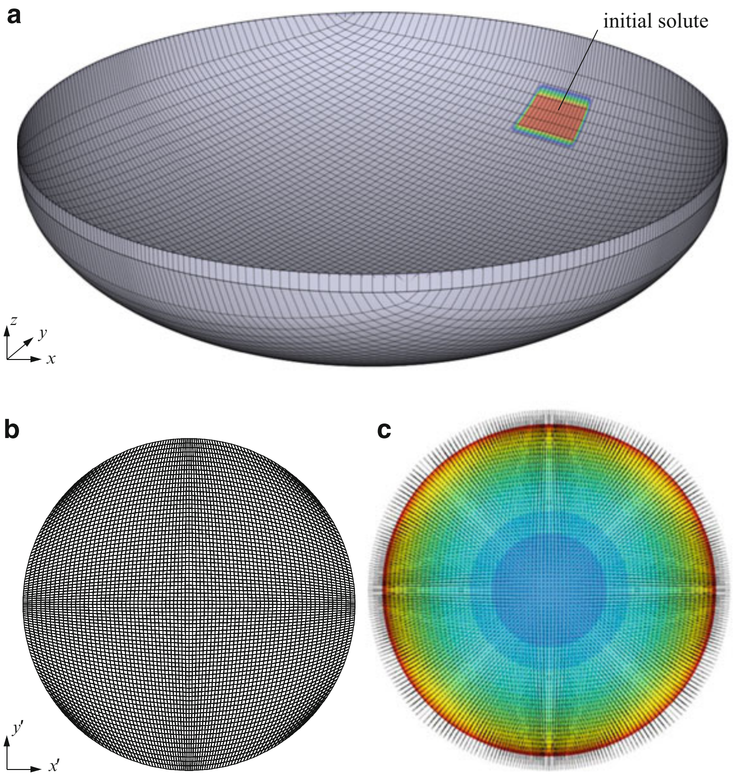


Fig. 11.14 Transformation of an idealized 3D hemispherical basin geometry into a 2D projected domain: (**a**) 3D geometry in global coordinates (vertical exaggeration 1:2), (**b**) 2D projected 'horizontal' mesh to be solved in local coordinates and (**c**) plot of projected gravity components in the $x' - y'$−plane

11.9.4 Limitations

This approach is applicable to relatively thin aquifers in which flow and density effects in the z'−direction perpendicular to layer-oriented principal directions are negligible. This can often be assumed for aquifer layers having a small slope or low curvature in their elevations. Furthermore, the solute (or heat) must be assumed invariable over the aquifer thickness (that means along z'). If the aquifer slope is becoming larger and the density effects are increasing, the variable-density flow process modeled by such a projected gravity field must be more and more inaccurate due to the fact that the approach suppresses vertical velocities even though becoming important. The accuracy particularly deteriorates with the increasing slope of the layer for free convection at a high density contrast (high Rayleigh number) when the solute (or heat) movement is fully gravity-driven. For mixed convection problems

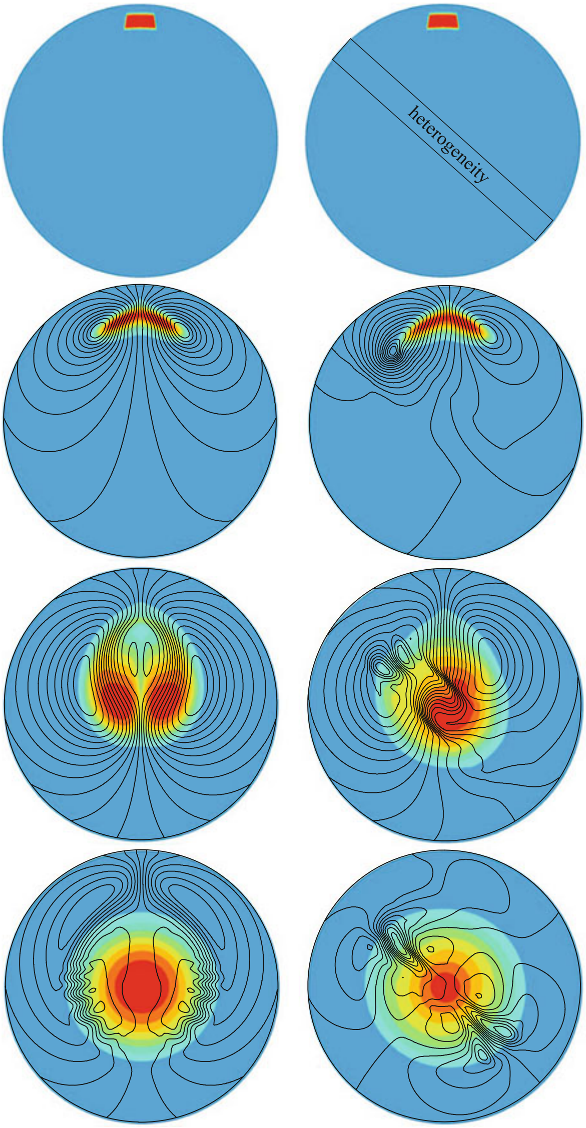


Fig. 11.15 Sinking down of a brine into a hemispherical basin in time, fringed solute distribution and contoured streamline pattern: (*left column*) homogeneous transmissivity, (*right column*) heterogeneous transmissivity distribution

(combined with a forced flow dynamics induced, for instance, by pumping), however, a larger slope in the geometry can often be tolerated. Generally, it is not possible to fix limits in form of critical slopes and curvature because it is widely dependent on the actual problem.

11.10 Non-Fickian Dispersion in Variable-Density Flow

11.10.1 Nonlinear Dispersion at High-Concentration Gradients

In modeling variable-density flow and mass transport problems an increasing interest has been cases where high-concentration (HC) differences in a 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. However, in 1D laboratory experiments [232, 464] with HC 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 saltwater 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 and Leijnse [232] and Hassanizadeh [225] 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. Experiments have confirmed these theoretical findings [464]. Furthermore, from the theoretical point of view the non-Fickian dispersion is consistent with the classic approach and theoretically well founded.

11.10.2 Extended Equations

In (11.1) only the standard linear Fick's law (3.183) has been incorporated into the species mass conservation equation. To extend the formulation to the non-Fickian law (3.187) we use the governing balance and phenomenological equations for species k as listed in Table 3.7:

$$\varepsilon s \acute{\Re}_k \frac{\partial C_k}{\partial t} + \boldsymbol{q} \cdot \nabla C_k + \nabla \cdot \boldsymbol{j}_k + (\varepsilon s \vartheta_k \Re_k + Q) C_k = \tilde{R}_k \tag{11.87}$$

or

$$\varepsilon s \acute{\Re}_k \frac{\partial C_k}{\partial t} + \boldsymbol{q} \cdot \nabla C_k - \nabla \cdot \Big(\frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k\| + 1} \cdot \nabla C_k \Big) + (\varepsilon s \vartheta_k \Re_k + Q) C_k = \tilde{R}_k \tag{11.88}$$

with

$$\boldsymbol{j}_k(\Im_H \|\boldsymbol{j}_k\| + 1) = -\boldsymbol{D}_k \cdot \nabla C_k \tag{11.89}$$

written for the convective form of the transport equation, where $(k = 1, \ldots, N)$, $\Im_H$ represents the additional high-concentration (HC) dispersion coefficient (possibly species-dependent in addition) required for the non-Fickian law and $\boldsymbol{D}_k$ is the known Bear-Scheidegger dispersion tensor (11.2) with longitudinal and transverse dispersivities, β_L and β_T, respectively, considered to be (constant) properties of the porous medium and independent of the fluid properties and transport process.

HC-gradient experiments [232,464] have shown that the nonlinear dispersion law (11.89) gives very good fits to measured breakthrough curves. It is found that the HC dispersion coefficient $\Im_H$ varies inversely with the flow velocity $\boldsymbol{q}$. Schotting et al. [464] have summarized their fitted experiments in the following approximate expression for $\Im_H = \Im_H(\boldsymbol{q})$ as

$$\Im_H(\boldsymbol{q}) = \frac{0.0125}{\|\boldsymbol{q}\|^{1.76}} \; [\mathrm{s\,m^2\,kg^{-1}}] \quad \text{for} \quad 9 \cdot 10^{-5} < \|\boldsymbol{q}\| < 3 \cdot 10^{-3} \; [\mathrm{m\,s^{-1}}] \tag{11.90}$$

11.10.3 Numerical Solution for Nonlinear Dispersion

The numerical solution of the governing balance equation (11.88) with the nonlinear dispersion law (11.89) requires a specific iterative treatment. The finite element formulations for the mass transport equations given in the preceding sections can be easily adapted to the non-Fickian law if we replace the linear tensor of hydrodynamic dispersion $\boldsymbol{D}_k$ by a nonlinear (extended) tensor of hydrodynamic dispersion $\boldsymbol{D}_k^\star$ in the form

$$\boldsymbol{D}_k \to \boldsymbol{D}_k^\star = \frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k\| + 1} \tag{11.91}$$

A recursive scheme is preferred which is performed by the following iteration procedure:

$$\begin{array}{lll}
\text{0. initial} & & \boldsymbol{j}_k^0 = \boldsymbol{0} \\
\text{1. step} & \boldsymbol{D}_k^\star = \frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k^0\|+1} & \boldsymbol{j}_k^1 = -\frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k^0\|+1} \cdot \nabla C_k^1 \\
\text{2. step} & \boldsymbol{D}_k^\star = \frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k^1\|+1} & \boldsymbol{j}_k^2 = -\frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k^1\|+1} \cdot \nabla C_k^2 \\
\vdots & \vdots & \vdots \\
\tau\text{. step} & \boldsymbol{D}_k^\star = \frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k^{\tau-1}\|+1} & \boldsymbol{j}_k^\tau = -\frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_k^{\tau-1}\|+1} \cdot \nabla C_k^\tau
\end{array} \tag{11.92}$$

where τ represents the iteration counter. The iteration (11.92) is done at each time step in dependence on the selected time stepping strategy: (1) For fixed (predefined) time steps it is iterated at each time plane. The procedure is terminated if the convergence criterion is satisfied. (2) For the GLS adaptive predictor-corrector time integration schemes (Sect. 11.6.4) the nonlinear solution is fully controlled by the time step itself, where the nonlinear dispersion tensor $\boldsymbol{D}_k^\star$ is linearized in time according to

$$\text{at new time plane } n+1\text{:} \quad \boldsymbol{D}_k^\star = \frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_{k,n+1}\| + 1} \quad \boldsymbol{j}_{k,n+1} = -\frac{\boldsymbol{D}_k}{\Im_H \|\boldsymbol{j}_{k,n}\| + 1} \cdot \nabla C_{k,n+1} \tag{11.93}$$

where the non-Fickian dispersive mass flux $\boldsymbol{j}_{k,n+1}$ at the new time plane $n+1$ is evaluated by using the non-Fickian dispersive mass flux $\boldsymbol{j}_{k,n}$ from the previous time plane n. At initial time the procedure is started with the standard linear Fick's law.

11.11 Benchmarks and Examples

As indicated in Sect. 1.2.2 the accuracy and reliability of models and algorithms have to be proved by procedures of verification, benchmarking and validation. The traditional verification procedure by use of analytical solutions is not generally applicable due to the nonlinear nature of variable-density problems. In general, there are no exact solutions for this problem class, except for a rather limited number of analytical and semianalytical solutions for specific cases [78, 107, 242, 470, 492]. As a consequence, modelers must rely on benchmark tests, which thus obtain a key role in proving variable-density flow models and simulation codes. Benchmarking covers asymptotic and mesh convergence tests, as well as comparative studies between different numerical solutions (mainly obtained with different simulators).

What are the characteristics of a valuable benchmark for variable density problems?

- It should have a real, practically and/or physically relevant background.
- It should be mathematically correct, definite and well-posed.
- Benchmark solutions should be predictable (nonrandom), both in the physical and mathematical sense.
- Ideally, the benchmark should have a physical model equivalent, for which qualified laboratory data are available. Those measurements can form reference solutions for a comparative analysis.

In this context, numerical solutions are required to understand the physical process and its causal dependencies. This is what we define as a *physical benchmark* that provides a physically reliable basis for further comparisons [138].

Physical benchmarks are frequently based on Hele-Shaw cell experiments, e.g., [97, 98, 153, 154, 481, 571, 572]. A Hele-Shaw cell provides an analog of flow in a porous medium to some extent, in that the equations that characterize flow in a Hele-Shaw cell (to a good approximation) are the same as the equations that characterize flow in a porous medium. However, the equations that characterize transport of salinity and heat in a Hele-Shaw cell are not quite the same as those that characterize transport of salinity and heat in a porous medium. For instance, dispersion and instabilities associated with 3D disturbances are different in a Hele-Shaw cell [325].

11.11.1 Hydrostatic Test

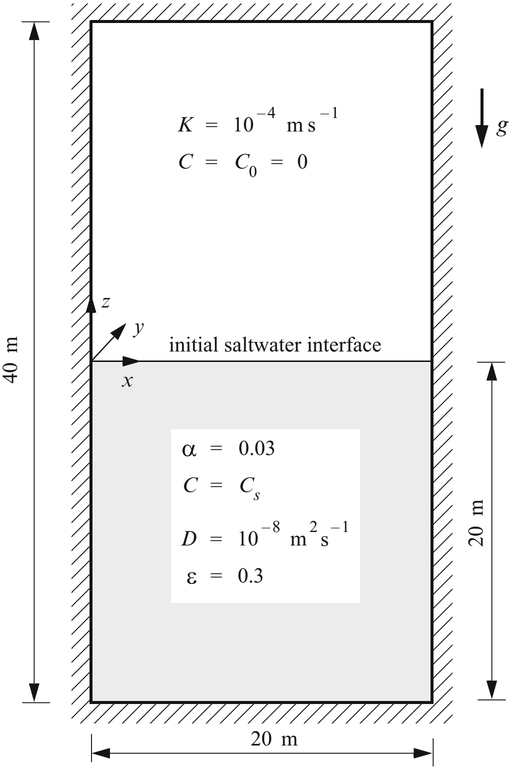
This type of benchmark is quite simple, but very instructive. It is a test of the velocity consistency (11.60) under hydrostatic and sharp density transition conditions as originally proposed by Voss and Souza [552]. Consider a rectangular closed domain as shown in Fig. 11.16. Initially, a stable saltwater layer with a salinity of $C = C_s$ exists below freshwater with $C = C_0 = 0$, separated by a sharp horizontal interface in the middle of the domain. (Note that the salinity C concerns a single-species concentration, where for the sake of simplicity we drop the species index k.) The boundary of the domain is impervious with respect to both the flow and the mass. The fluid density contrast (solutal expansion coefficient) α is defined by (3.275)

$$\alpha = \frac{\rho_s - \rho_0}{\rho_0} \quad \text{with} \quad \rho_s = \rho(C_s), \quad \rho_0 = \rho(C_0) \tag{11.94}$$

The $\alpha-$coefficient is to be varied in the numerical study whereby the density coupling strengthens with increasing α. We illustrate the results for an $\alpha-$value of 0.03.

The problem is hydrostatic at 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 the saltwater mixes (linear Fickian law is assumed) and the initially sharp saltwater interface (narrow transition zone) spreads in time. This process must be independent of the density. 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 $\alpha = 0$, based on a fine temporal and spatial discretization. We have simulated the density-dependent problem for 2D and 3D meshes by using the different velocity approximations (FKA and local smoothing technique), as discussed in Sects. 11.7.3 and 8.19.1.2. The findings are similar to those depicted in Fig. 11.17 for a 2D mesh consisting of 32 × 64 linear quadrilateral elements.

The local smoothing method causes an artificially increased spreading of the salinity much like numerical dispersion. This is caused by spurious local velocities at the interface nodes, that locally violate the consistency requirement. The smearing in the density profile increases, if dispersion ($\beta_L = 5\,\text{m}$, $\beta_T = 0.5\,\text{m}$) is additionally taken into account. In contrast, the consistent velocity approximation by the FKA avoids any spuriousness in the velocity field. These results agree very well with the reference solution, independent of dispersion effects.



$K = 10^{-4}\ \mathrm{m\,s^{-1}}$
$C = C_0 = 0$
g
z
y
x
initial saltwater interface
40 m
$\alpha = 0.03$
$C = C_s$
$D = 10^{-8}\ \mathrm{m^2 s^{-1}}$
$\varepsilon = 0.3$
20 m
20 m

Fig. 11.16 Cross-sectional view of the initially stratified (isothermal, single-species) saltwater below freshwater problem in a closed box of a saturated porous medium (Modified from [138])

This benchmark reveals the weakness of the smoothing methods, which normally work satisfactorily. However, these techniques cannot guarantee the local consistency in the velocity field for problems involving sharp transition zones in the density contrast and, therefore, advanced evaluation techniques such as the FKA are to be preferred. This benchmark can be extended by imposing a horizontal uniform flow at hydrostatic conditions. In such a test a density profile should not be smeared if transverse dispersivity and diffusion are set to zero.

11.11.2 Henry Problem

The Henry problem describes the advance of a saltwater front in a confined aquifer which was initially saturated with freshwater. Henry [242] developed a semianalytical solution technique for the steady-state case of this problem. Based on the OB approximation he derived analytical expressions for the streamfunction and the salt concentration in the form of Fourier series. The resulting algebraic equations for determining the coefficients of the Fourier series must be solved by numerical techniques. Using quite different approximation methods, a number of authors

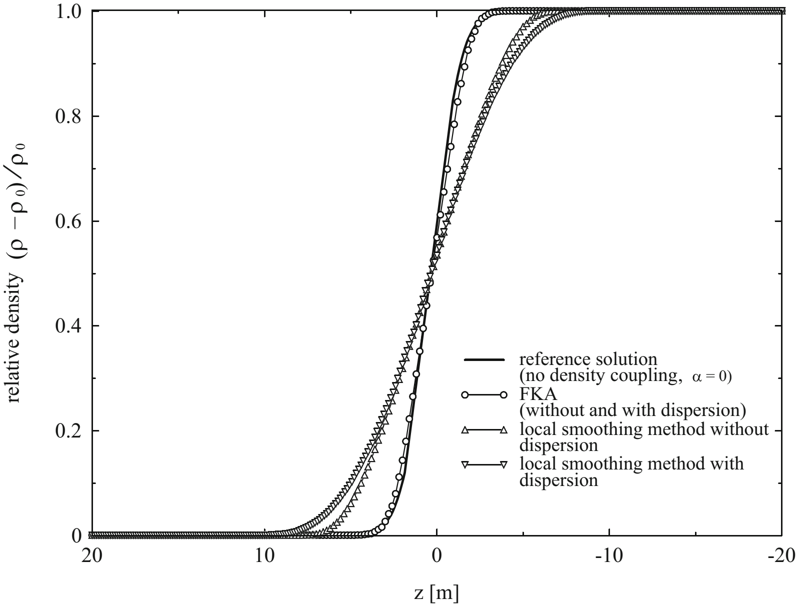


Fig. 11.17 Computed density profiles $[\rho(x, z) - \rho_0]/\rho_0$, $x = 10\,\text{m}$, $-20\,\text{m} \leq z \leq 20\,\text{m}$ at time $t = 10^3$ days for different solutions using quadrilateral elements: Reference solution is obtained without density effects for a fine vertical mesh; the other solutions are simulated on a uniform 32×64 mesh of linear quadrilateral elements (Modified from [138])

obtained similar results (e.g., Pinder and Cooper [420], Segol et al. [472], Desai and Contractor [122], Frind [174], Voss and Souza [552], Galeati et al. [180], Oldenburg and Pruess [398], Croucher and O'Sullivan [107], Kolditz et al. [318], Bués and Oltean [63]). The 'mystery' of Henry's solution is that no numerical model has been able to reproduce closely his semianalytical results [470] (cf. dashed line in Fig. 11.20). Only by modifying physical parameters in form of a reduced freshwater inflow or by a reduced dispersion the worthiness of the Henry problem could be improved [482, 588]. Nevertheless, as there exists no other non-numerical technique for this kind of nonlinear problem, Henry's solution has become one of the standard tests of variable-density groundwater models. The idealized aquifer for the simulation of Henry's problem is shown in Fig. 11.18. The BC's for flow consist of impermeable borders along the top and the bottom. Hydrostatic pressure is assumed along the vertical boundary of the sea side, which leads to a depth-variable BC for the hydaulic head h such as derived and discussed in Appendix L. The aquifer is charged with freshwater at a constant flux from the left side. At the inland side, the concentration is zero, which corresponds to a freshwater condition. At the coastal side the concentration of seawater is imposed. Instead of velocity-dependent dispersion a correspondingly large diffusivity was used by Henry [242]

Fig. 11.18 Definition of the Henry problem

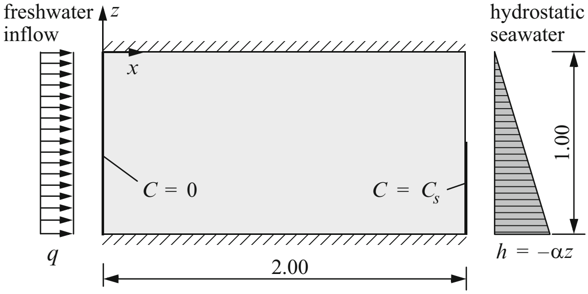


Table 11.4 Parameters and conditions used for the Henry problem

Quantity	Symbol	Value	Unit
Cell measure (length; height)		2; 1	m
Isotropic hydraulic conductivity	K	10^{-2}	$\mathrm{m\,s^{-1}}$
Specific storage coefficient	S_o	0	$\mathrm{m^{-1}}$
Specific solutal expansion coefficient	α	0.025	1
Porosity	ε	0.35	1
Molecular diffusion coefficient	D	$6.6{\cdot}10^{-6}$	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	0	m
Transverse dispersivity	β_T	0	m
Maximum concentration (salinity)	C_s	35	$\mathrm{g\,l^{-1}}$
Flow BC's			
Neumann-type BC at left side (inland)	q_h	−5.7024	$\mathrm{m\,d^{-1}}$
Hydrostatic head at right side (coast)	$h_D(z)$	$-\alpha z$	m
Mass IC and BC's			
Initial condition (IC) of salinity	C_0	0	$\mathrm{g\,l^{-1}}$
Dirichlet-type BC at left side (inland)	C_D	0	$\mathrm{g\,l^{-1}}$
Variants of Dirichlet-type BC's of salinity at right side (coast):			
Variant 1 – unconstrained salinity BC	$C_D = C_s \quad (-1 \le z \le -0.5)$		$\mathrm{g\,l^{-1}}$
Variant 2 – constrained salinity BC	$\begin{cases} C_D = C_s \\ Q_{n_C} > 0 \end{cases} \quad (-1 \le z \le 0)$		$\mathrm{g\,l^{-1}}$ $\mathrm{g\,d^{-1}}$
FEM			
2D mesh of 100 × 50 linear quadrilateral elements, GFEM (no upwind), OB approximation			
Initial time step size	Δt_0	10^{-4}	d
RMS error tolerance (AB/TR)	ϵ	10^{-3}	1
Simulation time period (reaching steady-state)	t_{end}	1	d

in order to allow a semi-analytical solution. The simulation parameters for the Henry problem are given in Table 11.4.

The steady-state flow pattern and the concentration distribution derived by Henry [242] are shown in Fig. 11.19. Additionally, the sharp-interface solution (Ghyben-Herzberg relation) is superposed on Henry's isochlors. Figure 11.20 summarizes some former findings for the Henry problem obtained by several authors, who used quite different computation methods. Comparing these results,

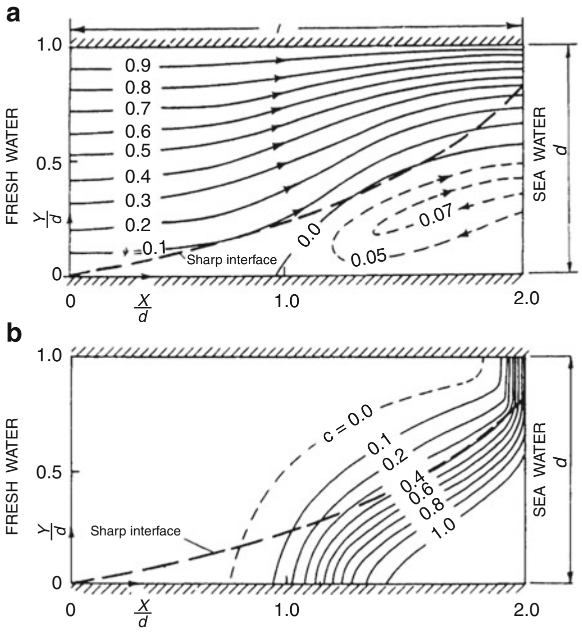


Fig. 11.19 (**a**) Streamlines and (**b**) concentration distribution C/C_s obtained by Henry from his semianalytical solution with Ghyben-Herzberg solution of sharp interface (From [242])

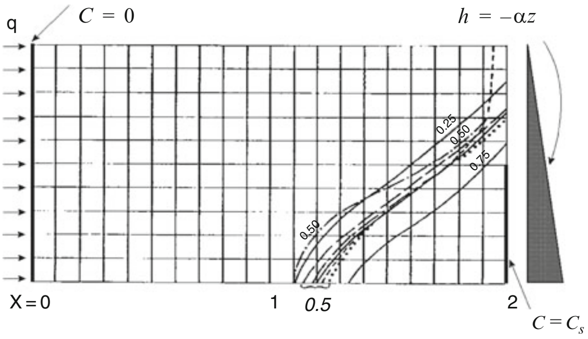


Fig. 11.20 Prior results by Henry [242] – *dashed line*, Pinder and Cooper [420] – *dashed-dotted line*, Segol et al. [472] – *dotted line*, Desai and Contractor [122] – *long-dashed line*, Frind [174] – *short-long-dashed line*, and Voss and Souza [552] – *solid line*; positions of the 25, 50, and 75 % isochlors of the steady-state solution

it has to be kept in mind that slightly different parameter values were chosen by the authors. There have been some discrepancies in the use of the diffusion coefficients. Further solutions and comparisons for the Henry problem have been presented by Kolditz et al. [318].

Present results are shown in Fig. 11.21 in form of computed isochlors at equilibrium (steady-state) simulated on a 2D uniform mesh consisting of 100 × 50 linear quadrilateral elements by using the conditions and parameters as listed in Table 11.4. The results are in good agreement with previous findings as displayed in Fig. 11.20. For the salinity BC at the coast side the two variants given in Table 11.4 are applied. In variant 1 the seawater concentration C_s is only imposed on the lower half of the vertical boundary, where it can be assumed that seawater enters the aquifer and no freshwater exits. Such a simplified BC is common and used

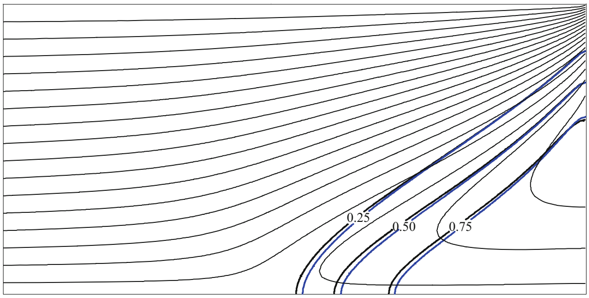


Fig. 11.21 Results for the Henry problem simulated on a 2D mesh consisting of 100 × 50 quadrilateral elements: computed streamlines and positions of the 0.25, 0.50 and 0.75 isochlors C/C_s in dependence on BC-variant 1 (*bold black*) and BC-variant 2 (*bold blue*) given in Table 11.4

in the most prior models. However, the vertical extent where seawater intrudes must be solution-dependent. A physically more realistic BC represents variant 2, where the complete boundary is imposed by the seawater concentration C_s, however, combined with a constraint allowing the salinity BC only when saltwater enters, while outflowing sections are switched to an open BC, $q_C \approx 0$, so that fresh and mixed water can freely pass the upper part of the boundary. Nevertheless, as seen in the results of Fig. 11.21 the differences between both BC variants are not remarkable for the Henry problem and the simpler unconstrained BC variant 1 appears suitable.

The Henry problem is often used in past as a benchmark for variable-density flow and single-species mass transport although it has some deficiencies. An unrealistically large amount of diffusion is introduced which results in a widely dispersed transition zone. It makes the solution smooth and easy. We conclude that the Henry problem is rather inappropriate for verifying purely density-driven flow situations.

11.11.3 Salt Dome Problem

This benchmark was proposed by the participants of the international HYDROCOIN project for the verification of groundwater models (Swedish Nuclear Power Inspectorate 1986). The test case is designed to model variable-density groundwater flow over a hypothetical salt dome, where the geometry is largely simplified. The geometry and BC's of the test problem are shown in Fig. 11.22. The cross section of the model extends horizontally 900 m and vertically 300 m. The aquifer is considered to be homogeneous, isotropic and saturated. The pressure varies linearly on the top of the aquifer. The other sides are impervious to flow. The single-species concentration (salinity) on the top is set to zero at the inflow domain. The middle section of the base represents the top of the salt dome with normalized mass concentration of solute $\hat{C} = C/C_s$ equal to unity. On all the remaining parts of the boundary, the normal concentration gradient was set to zero. The simulation parameters are listed in Table 11.5. The results are compared for the steady-state solutions.

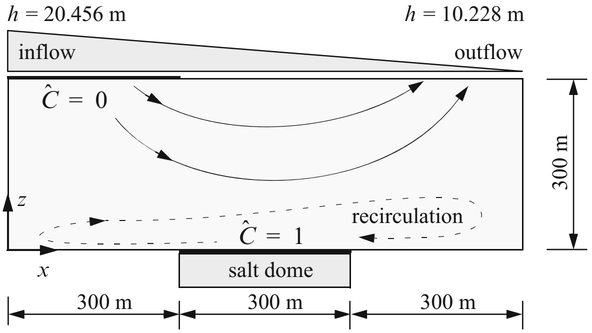


Fig. 11.22 Definition of the salt dome problem (HYDROCOIN Level 1 Case 5) – domain and BC's

Table 11.5 Parameters and conditions used for the salt dome problem

Quantity	Symbol	Value	Unit
Domain measure (length; height)		900; 300	m
Isotropic hydraulic conductivity	K	$1.09853 \cdot 10^{-5}$	$\mathrm{m\,s^{-1}}$
Specific storage coefficient	S_o	0	$\mathrm{m^{-1}}$
Specific solutal expansion coefficient	α	0.2036	1
Porosity	ε	0.2	1
Molecular diffusion coefficient	D	$1.39 \cdot 10^{-8}$	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	20	m
Transverse dispersivity	β_T	2	m
Flow BC			
Dirichlet-type BC at top	h_D	$20.456 - 0.011364x$	m
Normalized mass[a] *IC and BC's*			
Initial condition (IC) of salinity	$\hat{C}_0$	0	1
Dirichlet-type BC at top	$\hat{C}_D \quad (0 \leq x \leq 300)$	0	1
Dirichlet-type BC at bottom	$\hat{C}_D \quad (300 \leq x \leq 600)$	1	1
FEM			
2D mesh of 120 × 64 linear quadrilateral elements, GFEM (no upwind), OB approximation			
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period (reaching steady-state)	t_{end}	400	years

[a] Normalized salinity $\hat{C} = C/C_s$, where C_s occurs at the salt dome boundary

The salt dome was investigated by several authors (Herbert et al. [244], Leijnse [337], Oldenburg and Pruess [398], Oldenburg et al. [400], Johns and Rivera [290], Kolditz et al. [318], Konikow et al. [320], Holzbecher [254] and Younes et al. [583], among others). Kolditz et al. [318] obtained the same stratified system as Herbert et al. [244] already found. A freshwater region with higher velocities is observed in the upper part, where flow is driven by the superimposed pressure gradient on the top of the aquifer. There is a brine pool along the bottom, where flow with small velocities recirculates. The outflow of the saltwater is

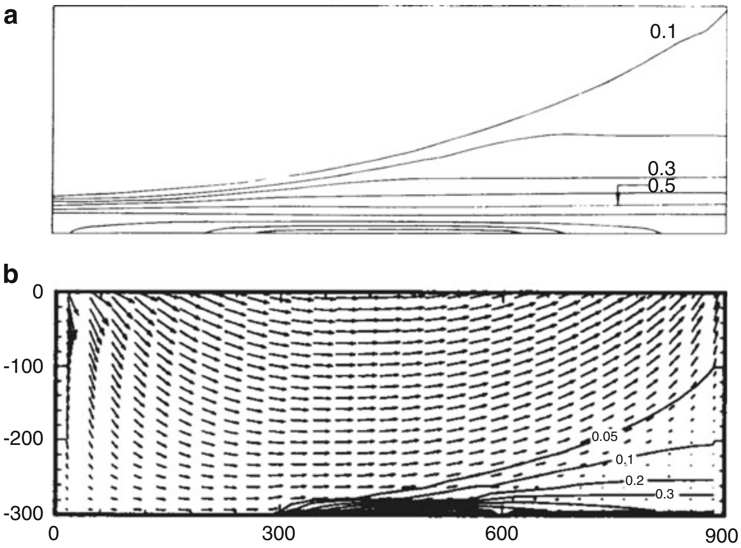


Fig. 11.23 Results for the salt dome problem: steady-state salinity contours simulated by (**a**) Herbert et al. [244] and (**b**) Oldenburg and Pruess [398]

focused on the upper right-hand corner of the domain. Johns and Rivera [290] has reproduced the prior results by Herbert et al. [244]. In contrast, different results have been presented by Oldenburg and Pruess [398], which has led to a broad discussion on the role of BC's and mechanical dispersion [254, 320, 400, 583]. The conflicting results are summarized in Fig. 11.23. Oldenburg and Pruess' solution is called 'fully swept-forward' pattern. It was shown [583] that both a swept-forward and a recirculation (Herbert-like) solution can be produced in dependence on the numerical representation of BC's (either salt concentration or salt mass flux) and the magnitudes of D, β_L and β_T.

The present simulations have been performed on a nonuniform mesh consisting of 120 × 64 linear quadrilateral elements having variable thicknesses in the vertical z−direction (smallest element thickness is 0.22 m at the bottom) by using the parameters listed in Table 11.5. The results obtained in form of salinity contours and streamlines shown in Fig. 11.24 confirm the recirculation-type pattern in a good agreement with previous findings [244, 254, 290, 318, 320, 583].

In deep aquifers the saltwater upconing process is subject to the buoyancy influences by a thermal gradient. Diersch and Kolditz [137] studied a thermohaline extension of the salt dome problem, where in addition a temperature difference between bottom and top boundary is applied (similar to the HRL problem of Sect. 11.5.1). Simulated results of the salt dome problem at a time of 100 years for different Turner numbers Tu, defined by (11.29), are shown in Fig. 11.25, demonstrating that the temperature effect on the saltwater distribution remains negligible or small if compared with the single-diffusive results at higher Turner

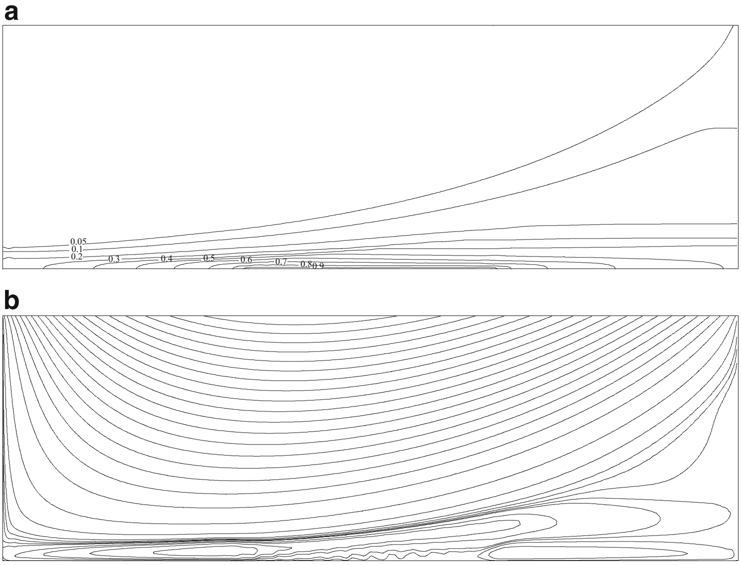


Fig. 11.24 FEFLOW results of the steady-state salt dome problem: (**a**) salinity contours $\hat{C} = (0.05, 0.1, 0.2, \ldots, 1.0)$ and (**b**) streamlines

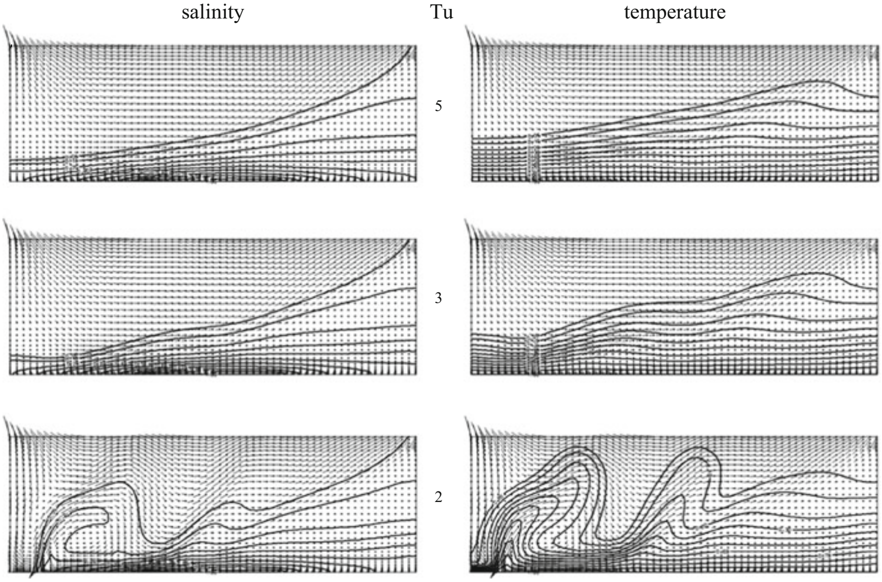


Fig. 11.25 FEFLOW results of the extended thermohaline salt dome problem: computed salinity and temperature distributions at 100 years for different Turner numbers Tu (Modified from [137])

numbers Tu. However, as seen for Tu = 2, if the Turner number becomes smaller, vigorous temperature influences on the brine pattern result in form of a ‘wavy’ salinity field caused by the thermal buoyancy. The ‘wavy’ salinity characteristics are triggered in front of the salt wedge by thermally driven eddies. As expected, this leads to an increased saltwater effluent on top of the aquifer. Note that a buoyancy ratio of Tu = 2 implies large temperature difference for a high-concentration brine and represents an extreme situation.

11.11.4 Elder Problem

The Elder problem serves as an example of free convection phenomena, where the bulk fluid flow is driven purely by fluid density differences. Elder [153, 154] presented experimental and numerical studies concerning the thermal convection produced, by heating a part of the base of a porous layer. The original experiment, which was performed in a Hele-Shaw cell, was called the ‘short-heater problem’. Elder conducted these studies mainly to verify the finite difference model he used for the 2D numerical analysis of thermal-driven convection. Furthermore, he suggested criteria for preventing numerical instabilities. Since then, Elder’s short-heater problem become a very popular and often stressed benchmark problem in the water resources literature. It is rich in physical and numerical implications, and its cellular flow characteristic is fascinating. The Elder problem has been modified, extended, and remains a topic of, sometimes controversial, discussion.

Diersch [130] and later Voss and Souza [552] transformed the thermal Elder problem into a solute-analogous convection problem, where heavy saltwater is placed on top. Voss and Souza ‘blew up’ the geometry so that Elder’s problem can now be deemed a large-scale, density-driven saltwater intrusion process in a cross-sectional aquifer schematization. The original Elder problem of the thermal convection in a Hele-Shaw cell and the solute-analogous convection problem are mathematically equivalent (via the Rayleigh number). However, we note that the problem in this formulation is completely imaginary and hypothetical. Dispersion, which can play a very important role in a real aquifer, is not included. Usually, this saline analog is also termed the Elder problem. The simulation parameters and BC’s for the saline Elder problem are summarized in Table 11.6 and Fig. 11.26.

Neither exact solutions nor qualified measurements (!) exist for the Elder problem. Consequently, the (currently) only way to compare is with numerical solutions. However, strong discretization effects were observed by using different meshes. Mesh convergence studies were conducted [138, 178, 288, 318, 401], where the meshes were consecutively refined until a supposed mesh convergence was achieved. Both the OB and the EOB approximation were studied [288, 318]. The most important outcome of these studies was, that for some meshes there could be a central upwelling flow while a central downwelling flow could be found for finer meshes, as exemplified in Fig. 11.27 for the 20 year evolution of the free convection. These observations gave rise to various numerical studies by

Table 11.6 Parameters and conditions used for the (saline) Elder problem

Quantity	Symbol	Value	Unit
Domain measure (length; height)		600; 150	m
Isotropic hydraulic conductivity	K	$4.754 \cdot 10^{-6}$	$\mathrm{m\,s^{-1}}$
Specific storage coefficient	S_o	0	$\mathrm{m^{-1}}$
Specific solutal expansion coefficient	α	0.2	1
Porosity	ε	0.1	1
Molecular diffusion coefficient	D	$3.565 \cdot 10^{-6}$	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	0	m
Transverse dispersivity	β_T	0	m
Solutal Rayleigh number	Ra	400	1
Flow BC			
Dirichlet-type BC at left-upper corner	$h(x,z) = h_D(0,150)$	0	m
Dirichlet-type BC at right-upper corner	$h(x,z) = h_D(600,150)$	0	m
Normalized mass[a] *IC and BC's*			
Initial condition (IC) of salinity	$\hat{C}_0$	0	1
Dirichlet-type BC at top ($z = 150$ m)	$\hat{C}_D \quad (150 \leq x \leq 450)$	1	1
Dirichlet-type BC at bottom ($z = 0$ m)	$\hat{C}_D \quad (0 \leq x \leq 600)$	0	1
FEM			
2D quadrilateral meshes of different resolution, GFEM (no upwind), OB approximation			
Initial time step size	Δt_0	10^{-3}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	20	years

[a] Normalized salinity $\hat{C} = C/C_s$, where C_s occurs at the upper boundary

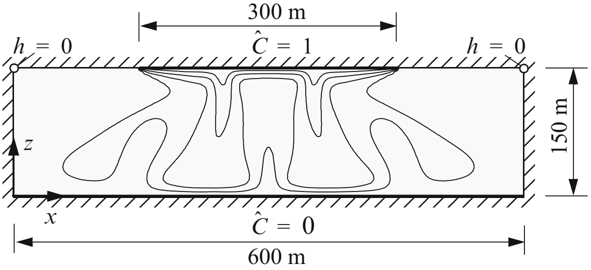


Fig. 11.26 Definition of the (saline) Elder problem – domain, BC's and salinity contours $\hat{C} = (0.2, 0.4, 0.6, 0.8)$ at $t = 10$ years

other authors [2, 70, 360, 410, 573]. The effects of heterogeneity in permeability distributions on Elder's convection process were studied by Prasad and Simmons [429] using a stochastic framework.

Mesh convergence studies concern a systematical mesh refinement for a mesh level $\ell = (0, 1, 2, \ldots)$. Using a uniform discretization by quadrilateral square elements the number of elements N_{E} and the number of nodes N_{P} for the whole and the half domain are given by

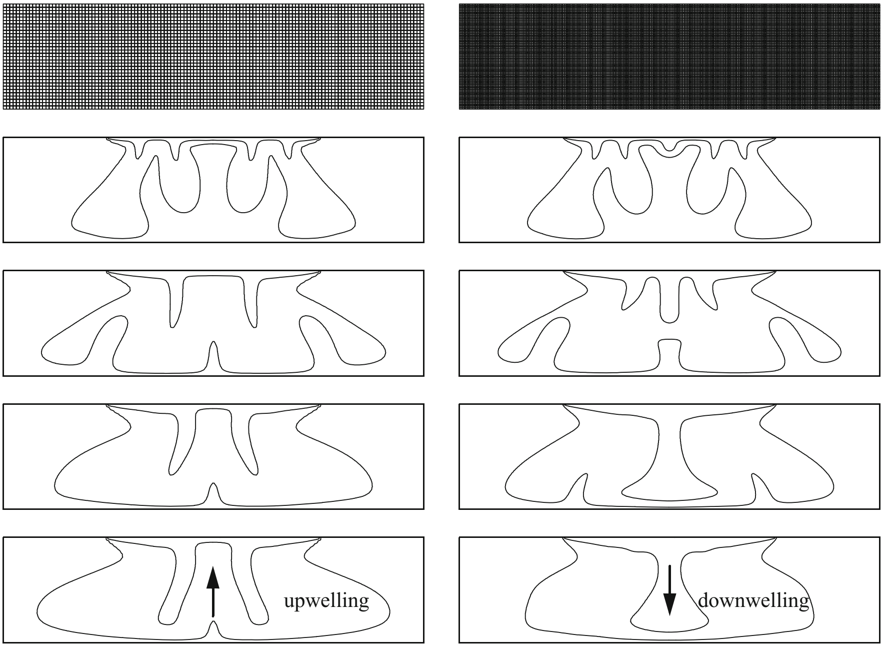


Fig. 11.27 Effect of spatial discretization on the computed salinity evolution at 5, 10, 15 and 20 year simulation time; positions of the 20 and 60 % isochlors: (*left column*) mesh with 4,257 nodes and 4,096 quadrilateral linear elements and (*right column*) mesh with 16,705 nodes and 16,384 quadrilateral linear elements

$$\begin{aligned} N_\mathrm{E} &= 2 \cdot N_{\mathrm{E}_2} & N_{\mathrm{E}_2} &= 2 \cdot 4^\ell \\ N_\mathrm{P} &= 2 \cdot N_{\mathrm{P}_2} - (2^\ell + 1) & N_{\mathrm{P}_2} &= 2 \cdot (2^\ell + 1)^2 - (2^\ell + 1) \end{aligned} \tag{11.95}$$

FEFLOW results for mesh levels ℓ up to 9 (this is a rather fine mesh with N_{P_2} equal to 525,825 for the half domain solved) are shown in Fig. 11.28, which confirm Frolkovič and De Schepper's observations [178] quite well. Note that in using a half domain model, an additional symmetry condition is imposed. Up to a mesh with resolution of $\ell = 5$, a central upwelling is seen which agrees with the previous findings reported by Oldenburg and Pruess [398], Kolditz et al. [318], Ackerer et al. [2], and Oltean and Bués [401]. From $\ell \geq 6$ onward the flow turns back to a downwelling pattern. This is shown in Fig. 11.28 for levels 6 and 9; levels 7 and 8 (not shown) are comparable with level 6. As revealed in the streamline pattern of Fig. 11.28, the flow behavior in the upper central location at a time between 2.5 and 5 years appears to be most critical for the further evolution of the convection process. Various eddies begin to mutually interact, to fuse and to disappear, and the local velocities control which rotation of the merged vortices finally prevails. At that location and time, the solution evolves either to an upwelling or to a downwelling flow regime. We can now summarize in Table 11.7 all of the findings relating to the

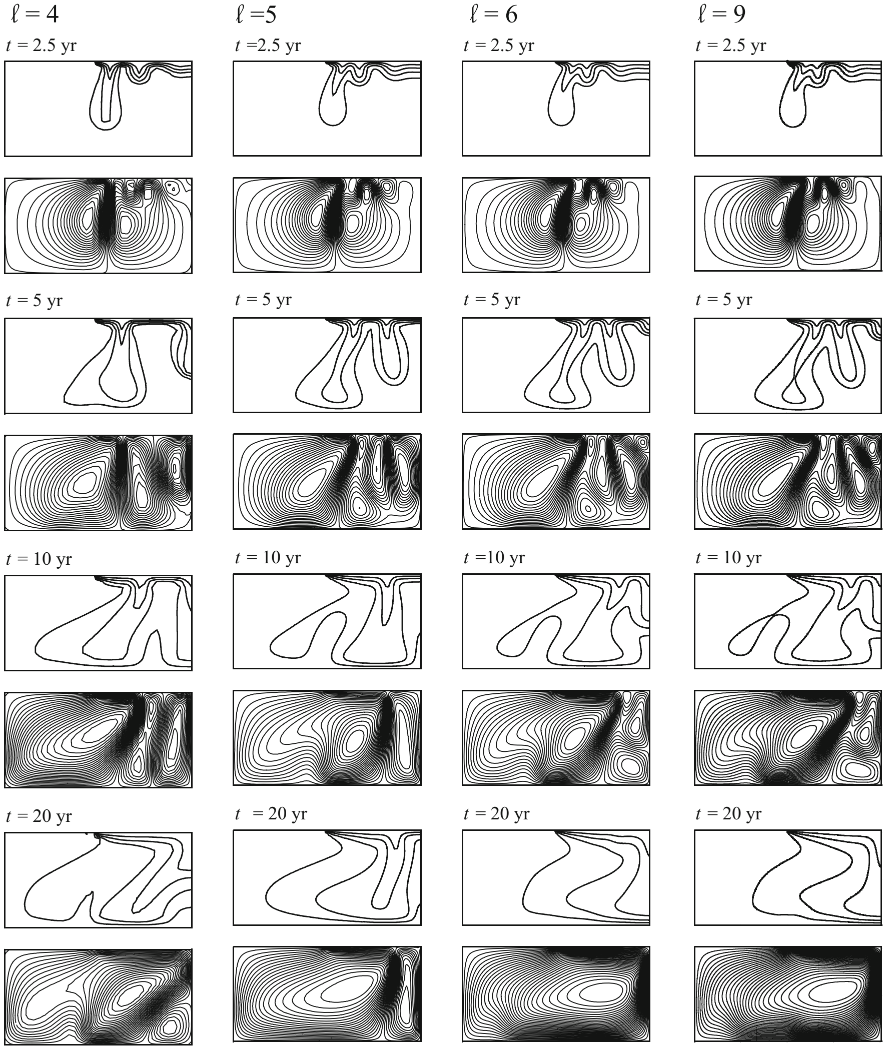


Fig. 11.28 Computed salinities (0.2, 0.4, 0.6 and 0.8 isolines) and streamline patterns for four times t (2.5, 5, 10, 20 years) and for four mesh levels ℓ (4,5,6,9): FEFLOW simulations using GFEM on quadrilateral uniform meshes of the half domain without perturbations and automatic AB/TR predictor-corrector time integration using FKA to compute consistent velocity fields (Modified from [138])

form of the flow direction in the central section, qualified by the degree of mesh refinement.

We evaluate the results in Table 11.7 only for their unperturbed solutions obtained on (mostly) uniform meshes. It is to be expected that uniform and aligned meshes with square elements are widely 'free' of perturbations, except

Table 11.7 Flow direction in the central section with respect to the mesh discretization: $\uparrow$ – upwelling, $\downarrow$ – downwelling (Modified from [401])

	Very			Very	Extremely
Discretization	coarse	Coarse	Fine	fine	fine
Number of unknowns[a]	<800	1,000 – 2,000	3,500 – 5,000	6,000 – 10^4	$1.5 \cdot 10^5 - 10^6$
Diersch [130]	$\downarrow$	–	–	–	–
Voss and Souza [552]	–	$\downarrow$	–	–	–
Oldenburg and Pruess [398]	–	$\downarrow$	$\uparrow$	$\uparrow$	–
Kolditz et al. [318]	–	$\downarrow$	$\uparrow$	$\uparrow$	–
Ackerer et al. [2]	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	–
Mazzia et al. [360]	–	$\downarrow$	–	–	–
Oltean and Bués [401]	$\uparrow\downarrow$	$\uparrow\downarrow$	$\uparrow$	$\uparrow$	–
Frolkovič and De Schepper [178][b]	–	$\downarrow$ ($\ell = 4$)	$\uparrow$ ($\ell = 5$)	–	$\uparrow$ ($\ell = 6$) $\downarrow$ ($\ell = 7$)
Diersch and Kolditz [138][c]	–	$\downarrow$ ($\ell = 4$)	$\uparrow$ ($\ell = 5$)	–	$\downarrow$ ($\ell = 6, 7, 8, 9$)
Johannsen [288]	–	–	$\uparrow\downarrow$ ($\ell = 5$)	–	$\uparrow\downarrow$ ($\ell = 6, 7, 8$)
Park and Aral [410]	–	$\downarrow$ ($\ell = 4$)	–	–	$\downarrow$ ($\ell = 6$)
Woods and Carey [573][d]	–	–	$\downarrow$ ($\ell = 5$)	–	$\uparrow$ ($\ell = 6, 7, 8$)

[a] Related to the whole domain

[b] Unperturbed solutions, uniform mesh of linear elements

[c] No upwind, AB/TR time integration, uniform linear quadrilaterals, unperturbed solutions

[d] Using quadratic elements, AB/TR time integration, uniform mesh

'noise' of numerical round-off and discretization errors. Furthermore, Frolkovič and De Schepper [178] also perturbed the problem by slightly modifying IC's for one of the vertices of the square element at the upper right corner of the half domain. For smaller times ($t < 4$ years) there were no remarkable changes in the solutions, but at later times the solution evolves in different directions. They observed three directional behaviors which convert to three different stationary solutions in form of a downwelling, an upwelling and a modified downwelling pattern (Fig. 11.30). All three solutions could be reproduced for various mesh levels. They concluded that nonunique stationary solutions exist for the Elder problem. It is suggested that unaligned (unstructured) meshes can cause a perturbation which, eventually, may determine the character of the numerical solution.

The convective behavior of the system can be well characterized by the vertical solute flux entering and leaving the convection cell on top and bottom, respectively, in form of the Sherwood number Sh (11.31), which can be easily computed by boundary flux evaluations via CBFM (cf. Sect. 8.19.2). For example, Fig. 11.29 exhibits the history of Sh for the Elder problem simulated for the mesh level $\ell = 6$. It indicates that the convection at $\text{Ra} = 400$ leads to an about 5.7 times larger mass throughput in approaching to steady state compared to the stationary pure diffusion at $\text{Ra} = 0$.

In a systematic bifurcation analysis based on FVM, Johannsen [288] also identified three stable steady-state solutions for the Elder problem at $\text{Ra} = 400$. In a highly accurate pseudospectral approach Frolkovič and De Schepper's as well as Johannsen's findings in form of the three stable steady-state solutions of the Elder problem as exhibited in Fig. 11.30 could be fully confirmed by Reeuwijk et al. [543],

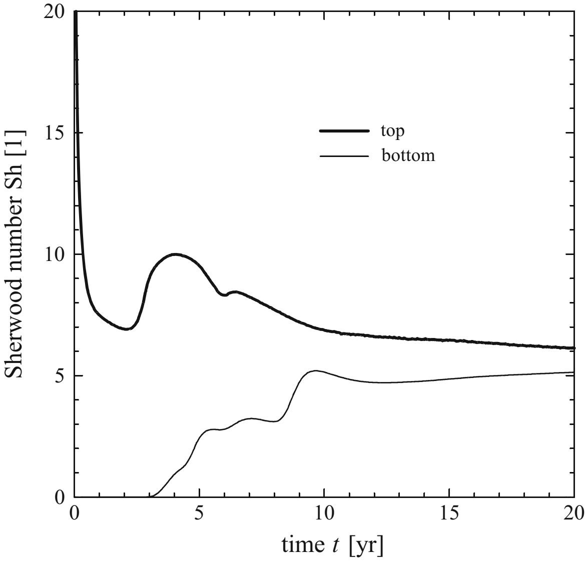


Fig. 11.29 History of vertical solute flux Sh computed at *top* and *bottom* boundary for mesh level $\ell = 6$

which were also shown in a perfect agreement with FEFLOW results. The three solutions are denoted by S_1, S_2 and S_3, where the subscript represents the number of downward plumes in the solution. An interesting outcome of Reeuwijk et al.'s analysis [543] is that the solutions S_1 and S_2 can be found quite easily by using IC's perturbed somehow, while solution S_3 can be achieved only for a small subset of perturbed IC's.

The nonuniqueness of the Elder problem could question its usability for benchmarking purposes. However, we have to take into account that for free convection problems at a sufficiently high Rayleigh number the existence of multiple solutions is the rule rather than the exception. It means for the Elder problem a benchmark test like this should reproduce at least one of the three solutions of Fig. 11.30, where solutions S_1 and S_2 are more likely.

Diersch and Kolditz [137] extended the originally 2D Elder problem to three dimensions for studying both the saline convection and the DDC processes. The 3D counterpart consists of a porous box with a square base ($600 \times 600\,\text{m}^2$) and height $H = 150\,\text{m}$. This box has the same cross-sections along the Cartesian axes as defined in Fig. 11.26 for the 2D sketch. Salinity is held constant, in an areal extent, on the top of the porous box. The parameters correspond to those given in Table 11.6. In Diersch and Kolditz's simulations a GFEM and an AB/TR time integration was used, where only the symmetric quarter of the domain is discretized by 48,000 linear hexahedral elements with 51,701 nodes (Fig. 11.31). In comparison with the mesh requirements we have seen for the 2D problem, such a 3D discretization is considered a 'moderate' resolution.

Fig. 11.30 The three stable steady-state solutions of the Elder problem: (**a**) S_1 – one downward plume with central downwelling, (**b**) S_2 – two downward plumes with central upwelling and (**c**) S_3 – three downward plumes with central downwelling (Modified from [543])

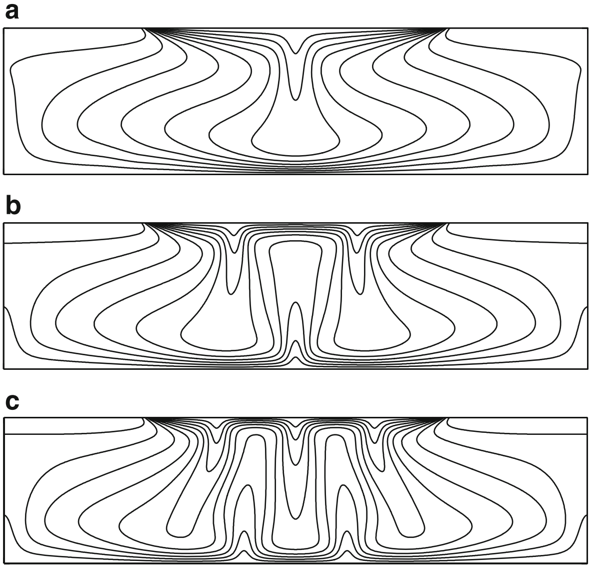
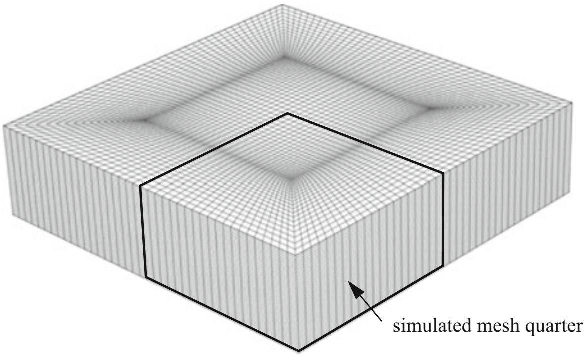


Fig. 11.31 Finite element mesh of the 3D Elder problem ($N_\mathrm{E} = 48{,}000$ linear hexahedra and $N_\mathrm{P} = 51{,}701$ to discretize the symmetric quarter of the domain)



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. 11.32 shows the evolution of salinity from different views. The 3D cut-away images (left column of Fig. 11.32) display the progressing fingering characteristics in the 3D space. Similar to the 2D case we find also an upwelling salinity pattern in the center of the box at the given time stages. The 3D influence becomes also apparent in the two horizontal views at an upper elevation of $0.9H$ (135 m) and the middle horizon of $0.5H$ (75 m) as shown in Fig. 11.32. 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

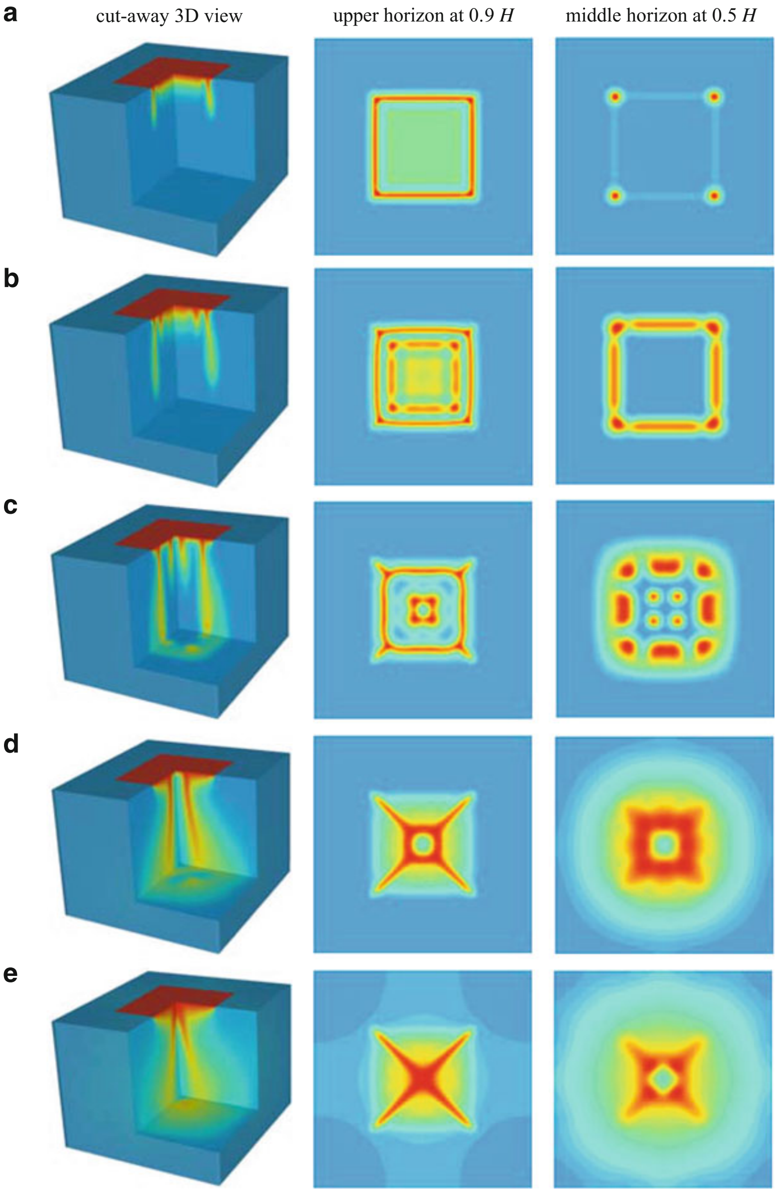


Fig. 11.32 Computed salinity patterns of the 3D Elder problem at times of (**a**) 1, (**b**) 2, (**c**) 4, (**d**) 10 and (**e**) 20 years (vertical exaggeration 2.6:1) (Modified from [137])

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.

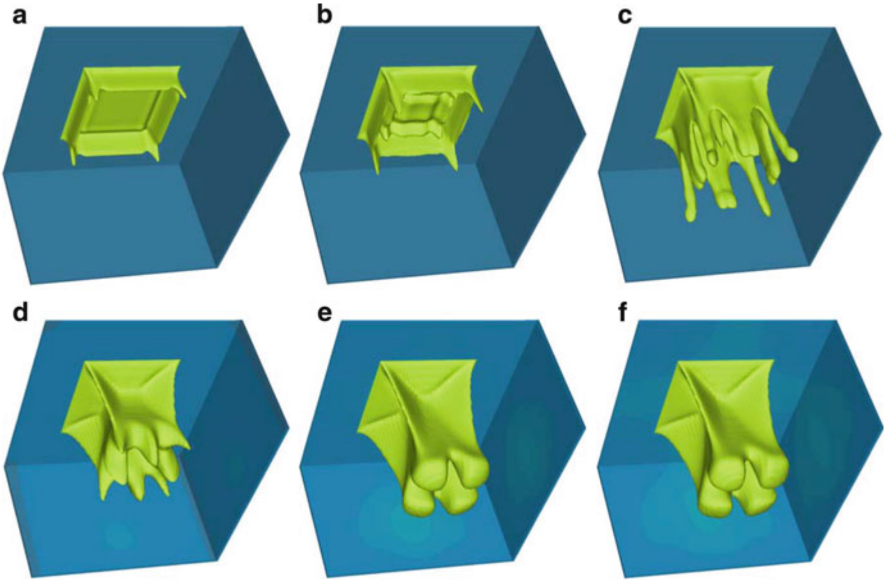


Fig. 11.33 Computed 3D isosurfaces of 50 % salinity for the 3D Elder problem (viewing into the *box* from *bottom* to *top*) at times of (**a**) 1, (**b**) 2, (**c**) 4, (**d**) 10, (**e**) 15 and (**f**) 20 years (vertical exaggeration 2.6:1) (Modified from [137])

An illustration of the pattern evolution in 3D space is given in Fig. 11.33 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.

The computed salinity patterns of the 3D Elder problem have been confirmed by the simulations performed by Mazzia and Putti [359] based on a mixed FEM approach and tetrahedral meshing. Results of the 3D Elder problem extended to a 3D thermohaline problem have been reported by Diersch and Kolditz [137, 138].

11.11.5 Salt Lake Problem

The salt lake problem as a test case for variable-density saturated flow and solute transport was introduced by Simmons et al. [479]. It represents a convection process below an evaporating salt lake, which can be observed in shallow playa groundwater systems (Fig. 11.34a). The evaporation process results in dense brine overlying less dense fluid leading to a downward convection of salt fingers. The numerical results modeled on a 2D schematization (Fig. 11.34b) were compared with those from a laboratory Hele-Shaw cell developed by Wooding et al. [571, 572]. Figure 11.34c gives the layout of the experimental Hele-Shaw cell. The tilted cell has a slope at an angle

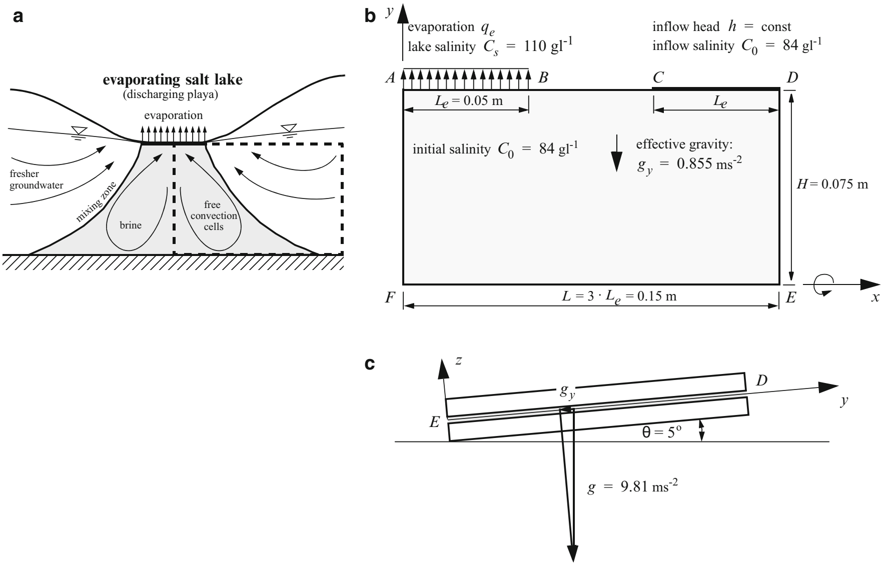


Fig. 11.34 Definition of the salt lake problem: (**a**) conceptual model of brine reflux in a shallow playa groundwater system (Modified from [390]), (**b**) idealized 2D model domain and (**c**) Hele-Shaw cell analog showing gravity component $g_y = g \sin \theta$ effective in the slit plane

of 5° to the horizontal. The model parameters are listed in Table 11.8. Unspecified BC's represent no-flow boundaries, at which natural BC's are imposed. Wooding et al. [571, 572] described the observations from their experiments as follows: '*At early times many small plumes grow from the evaporating boundary layer. These plumes descend under gravity and tended to coalesce to form larger-scale fingers. Differential growth and coalescence as seen from the Hele-Shaw cell results … are plausible mechanisms which allow for the growth of millimeter- or centimeter-scale wavelength, corresponding to wavelengths of meters or tens of meters or more which might be possible in nature. These larger plumes tended to maintain their identity during growth. By (dimensionless time)* $\hat{t} = 15.99$ *the leading plume had encountered the lower impermeable boundary and had begun to spread.*'

Simmons et al. [479] used two numerical codes for the analysis of the salt lake problem: the SUTRA simulator developed by Voss [550], and a streamfunction based finite difference model developed by Wooding et al. [571,572]. Simmons et al. [479] obtained a reasonable spatial and temporal agreement between the numerical and experimental outcomes. The criteria they choose for comparing the results were: (i) qualitative comparison of the fingering pattern and coalescence with time, (ii) examination of the effect of background advection on the movement of the finger sequence to the left of the cell, (iii) comparison of vertical growth rates of fingers as they move downwards, and (iv) representation of the entrainment of smaller fingers by the larger leading plume that originates at the boundary of the salt lake.

For the FEFLOW simulations of the salt lake problem several structured and unstructured meshes with different resolutions in combination with various

Table 11.8 Parameters and conditions used for the salt lake problem

Quantity	Symbol	Value	Unit
Cell measure (length; height)	L; H	0.15; 0.075	m
Cell evaporation length	L_e	$5 \cdot 10^{-2}$	m
Cell plate spacing	b	$2.1 \cdot 10^{-4}$	m
Cell angle to the horizontal	θ	5	°
Effective gravity	$g_y = g \sin\theta$	0.855	$\mathrm{m\,s^{-2}}$
Salinity at inflow and initial time	C_0	84	$\mathrm{kg\,m^{-3}}$
Salinity of the lake	C_s	110	$\mathrm{kg\,m^{-3}}$
Reference liquid density at inflow	$\rho_0 = \rho(C_0)$	$1.0646 \cdot 10^3$	$\mathrm{kg\,m^{-3}}$
Liquid density change [479]	$\frac{\partial\rho}{\partial C}$	0.780	1
Specific solutal expansion coefficient[a]	$\alpha = \frac{\rho(C_s)-\rho_0}{\rho_0}$	$1.9 \cdot 10^{-2}$	1
Dynamic viscosity of liquid	μ_0	$1.1 \cdot 10^{-3}$	$\mathrm{kg\,m^{-1}\,s^{-1}}$
Cell intrinsic permeability	$k = b^2/12$	$3.68 \cdot 10^{-9}$	$\mathrm{m^2}$
Isotropic hydraulic conductivity	$K = \frac{k\rho_0 g_y}{\mu_0}$	$3.045 \cdot 10^{-3}$	$\mathrm{m\,s^{-1}}$
Porosity	ε	1	1
Molecular diffusion coefficient	D	$9 \cdot 10^{-10}$	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	$9 \cdot 10^{-10}$	m
Transverse dispersivity	β_T	$9 \cdot 10^{-10}$	m
Evaporation rate	q_e	$1.03 \cdot 10^{-6}$	$\mathrm{m\,s^{-1}}$
Solutal Rayleigh number	$\mathrm{Ra} = \frac{\alpha K H}{\varepsilon D + \beta_T q_e}$	4,821	1
Flow BC			
Neumann-type BC at $\widehat{AB}$[b]	$q_h = q_e$	0.088992	$\mathrm{m\,d^{-1}}$
Dirichlet-type BC at $\widehat{CD}$[b]	h_D	0	m
Mass IC and BC's			
Initial condition (IC) of salinity	C_0	84	$\mathrm{kg\,m^{-3}}$
Dirichlet-type BC $\widehat{AB}$[b]	$C_D = C_s$	110	$\mathrm{kg\,m^{-3}}$
Dirichlet-type BC at $\widehat{CD}$[b]	$C_D = C_0$	84	$\mathrm{kg\,m^{-3}}$
FEM			
Nonuniform 2D meshes of different resolutions, GFEM and FU, OB approximation			
Initial time step size	Δt_0	10^{-7}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period[c]	$\hat{t}_{\mathrm{end}}$	18	1

[a] $\rho(C_s) = \rho_0 + \frac{\partial\rho}{\partial C}(C_s - C_0)$

[b] Defined in Fig. 11.34b

[c] Dimensionless time: $\hat{t} = \frac{\alpha K}{\varepsilon H} t$

numerical options have been used. The nearest similarity to the SUTRA findings using an identical spatial resolution (4,876 nodes and 4,725 elements) was obtained with an AB/TR time stepping scheme and a full upwind (FU) technique (Fig. 11.35). However, the salt fingers arrive at the cell bottom earlier than in the SUTRA simulations. Additionally, the development of smaller scale fingers at the salt lake boundary is suppressed. Refined meshes (we used up to 169,621 triangular elements with 85,401 nodes) seem to confirm the general features as upwelling-downwelling pattern and the formation of small fingers at the right lake boundary, but the evolution and number of intermediate convection cells is different (Fig. 11.36).

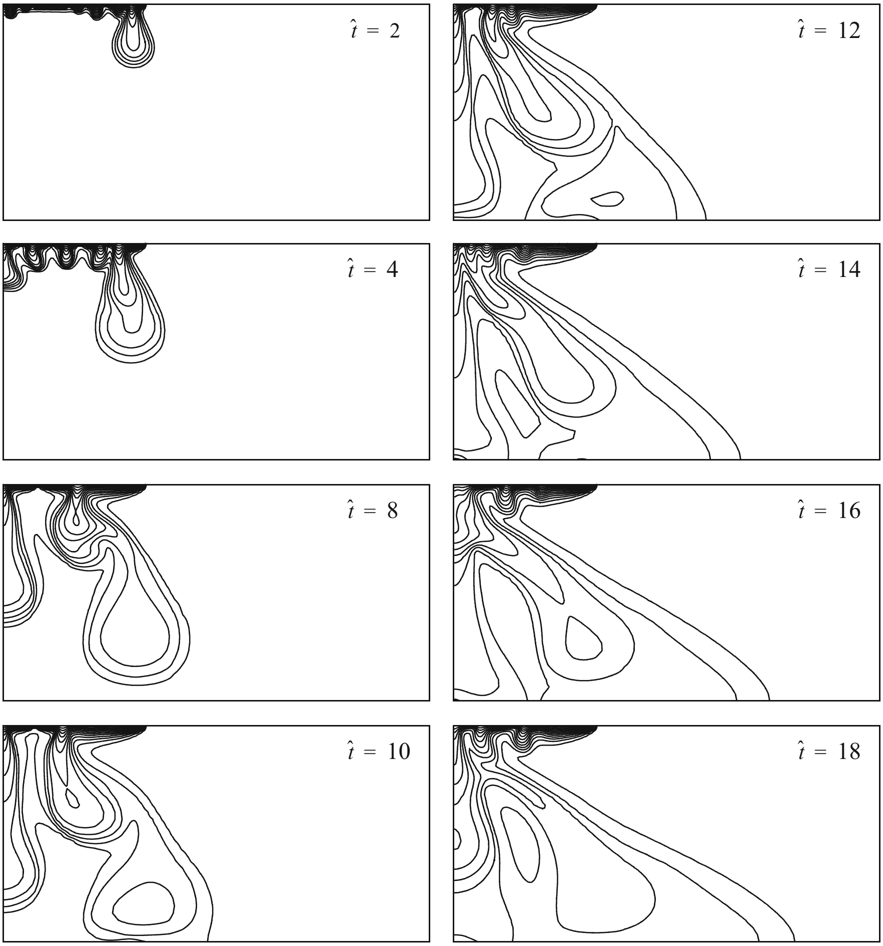


Fig. 11.35 Salinities C for different dimensionless times $\hat{t} = \frac{\alpha K}{\varepsilon H} t$. Contour interval is $2\,\mathrm{g\,l^{-1}}$. FEFLOW simulations on a coarse mesh (4,733 linear quadrilateral elements), full upwinding and AB/TR time stepping. Total number of adaptive times steps is 290 (Modified from [138])

These results seem to compare better to the experimental findings by Wooding et al. [571,572]. In a recent numerical study Wooding [570] found plume contours at early stages similar to the results of Fig. 11.36.

Mazzia et al. [360] attempted a mesh convergence study for the salt lake problem by using a mixed hybrid FEM. Three meshes were studied, with the finest one consisting of 40,000 triangles with 20,301 edges. Their results have a close similarity with our predictions shown in Fig. 11.36. In agreement with our observations they also found that the predicted finger patterns agree paradoxically much better with the laboratory observations, when simulated on coarse, instead on

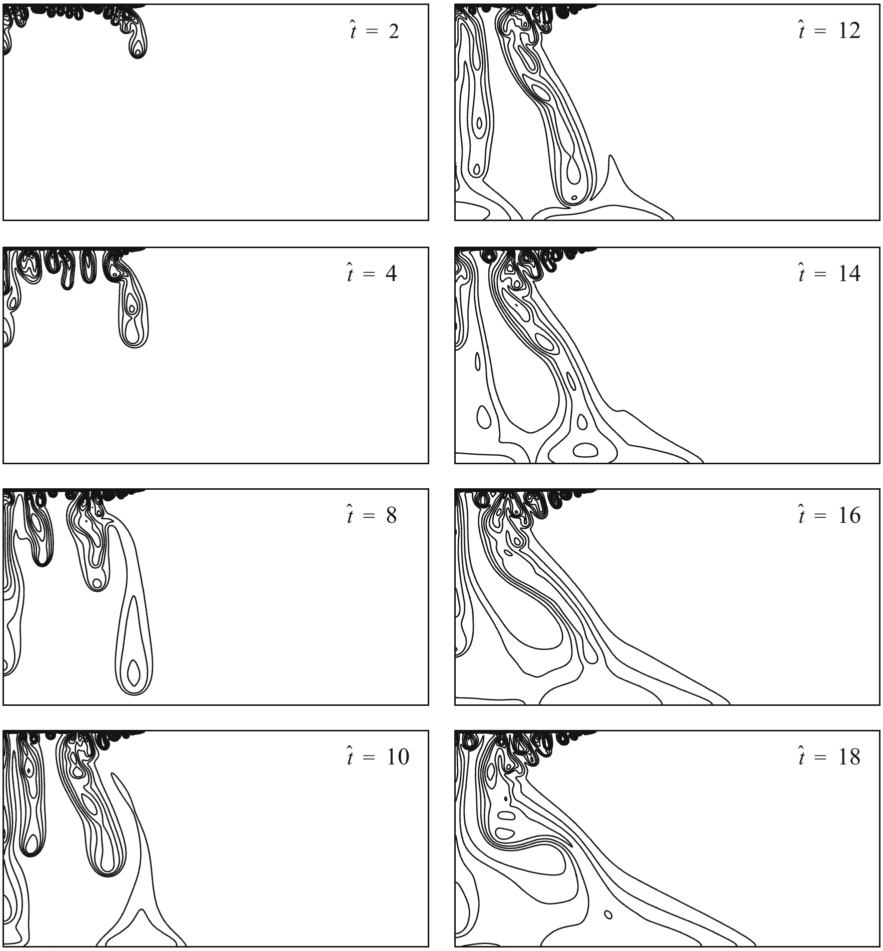


Fig. 11.36 Salinities C for different dimensionless times $\hat{t} = \frac{\alpha K}{\varepsilon H} t$. Contour interval is $2\,\mathrm{g\,l^{-1}}$. FEFLOW simulations on a fine mesh (169,621 linear triangular elements), GFEM (no upwind) and AB/TR time stepping. Total number of adaptive times steps is 2,700 (Modified from [138])

fine meshes. A mesh convergence was not achieved and they critically concluded that such an assessment is problematic.

It is obvious that the simulations significantly depend on the discretization and numerical features. We remark that the Rayleigh number of about 4,800 for the salt lake problem is more than 10 times larger than for the Elder problem, and we recall that a convection process with a Rayleigh number Ra $>$ 1,000 is in a range where different branches of flow regimes may exist. Accordingly, the main difficulty must be expected in the extremely dynamic behavior of the convection process, where physical perturbations caused by laboratory-scale heterogeneities that trigger instabilities must be mimicked in a numerical simulation.

Table 11.9 Simulation parameters for the HC displacement experiment

Quantity	Symbol	Value	Unit
Column length ($-4.0\,\mathrm{m} \leq z \leq 0.5\,\mathrm{m}$)[a]	L	4.5	m
Porosity	ε	0.2	1
Flow rate	q_o	$3.209 \cdot 10^{-5}$	$\mathrm{m\,s^{-1}}$
Reference salinity (freshwater)	C_0	0	$\mathrm{kg\,m^{-3}}$
Brine input	C_s	285.714	$\mathrm{kg\,m^{-3}}$
Molecular diffusion coefficient	D	0	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	1.0	m
HC dispersion coefficient	$\Im_H$	10^4	$\mathrm{m^2\,s\,kg^{-1}}$
Flow BC			
Neumann-type BC at top ($z = 0.5\,\mathrm{m}$)	$q_h = -q_o$	-2.772576	$\mathrm{m\,d^{-1}}$
Dirichlet-type BC at bottom ($z = -4.0\,\mathrm{m}$)	h_D	0	m
Mass (brine) IC and BC's			
Initial condition (IC)	C_0	$\begin{cases} C_s & \text{for } z \geq 0 \\ 0 & \text{for } z < 0 \end{cases}$	$\mathrm{kg\,m^{-3}}$
Dirichlet-type BC at top ($z = 0.5\,\mathrm{m}$)	$C_D = C_s$	285.714	$\mathrm{kg\,m^{-3}}$
Neumann-type BC at bottom ($z = -4.0\,\mathrm{m}$)	q_C	0	$\mathrm{kg\,m^{-2}\,s^{-1}}$
FEM			
Space increment	Δz	$5 \cdot 10^{-3}$	m
Initial time step size	Δt_0	10^{-7}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period[b]	$\hat{t}_{\mathrm{end}}$	1	1

[a] z directed upward

[b] Dimensionless time: $\hat{t} = \frac{q_o}{\varepsilon \beta_L} t$

11.11.6 High Concentration Flow Through a Column

High concentration-gradient (HC) experiments in a column of glass beads [232,464] have shown that the nonlinear dispersion law (11.89) gives very good agreements with measured breakthrough curves for brines. Schotting et al. [464] have derived analytical solutions in one dimension, which can be used to verify the approach for the non-Fickian dispersion law as described in Sect. 11.10. We consider the displacement of a high concentration through a column with constant properties. The parameters are summarized in Table 11.9.

On top of the column brine C_s starts entering the column with a uniform specific discharge q_o. A natural BC is imposed on the outflowing boundary at bottom. The column is discretized by 900 linear quadrilateral elements resulting in a spatial increment of $\Delta z = 0.005\,\mathrm{m}$. For the temporal approximation the AB/TR predictor-corrector time integration is used. It requires 144 time steps to simulate the displacement process for a dimensionless time $\hat{t} = \frac{q_o}{\varepsilon \beta_L} t$ up to $\hat{t} = 1.0$. The numerical results are in a very good agreement with the analytical results given by Schotting et al. [464] as shown in Fig. 11.37.

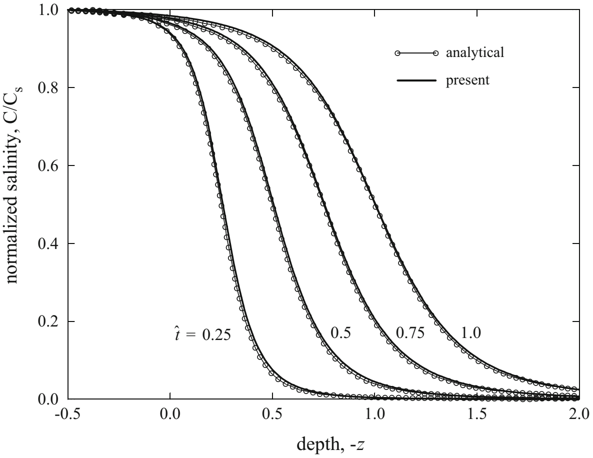


Fig. 11.37 Numerical density profiles simulated by FEFLOW at selected dimensionless times $\hat{t} = \frac{q_o}{\varepsilon \beta_L} t$ in comparison with the semi-explicit analytical solutions given by Schotting et al. [464] for a brine displacement in a column at nonlinear dispersion (Modified from [138])

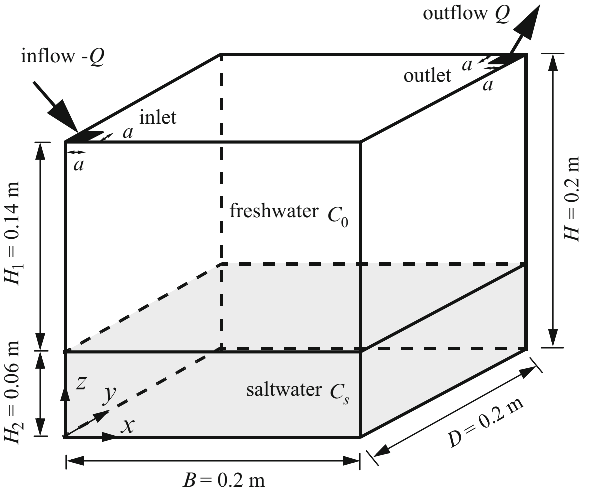


Fig. 11.38 Definition of the saltpool problem

11.11.7 Saltpool Problem

The saltpool problem was introduced by Oswald [404] and Oswald and Kinzelbach [405, 406]. It represents a 3D saltwater upconing process in a cubic box under the influence of density and hydrodynamic dispersion. A stable layering of saltwater below freshwater is considered in time for two cases: (1) low density case (1 % salt mass fraction) and (2) high density case (10 % salt mass fraction).

The experimental set-up consists of a cubic container covered by plexiglass walls and filled with dry silica glass beads (average diameter 1.2 mm). At the beginning of the experiment, saltwater is layered below freshwater, forming a horizontal narrow transition zone. Inflow and outflow were possible only via small holes in the corners of the test cube (Fig. 11.38). The box is recharged with freshwater through a single inflow opening at a constant rate Q. Water discharges through the outlet with a variable salinity. In the experiments, salinity breakthrough curves at the outflow opening were measured. The measured mixing concentration at the outflow is in fact very small, i.e., in the order of $\frac{1}{100}$ and $\frac{1}{1{,}000}$ related to the maximum salinity C_s for the low and high density cases, respectively. The position of the saltwater-freshwater interface was determined by use of the nuclear magnetic resonance (NMR) technique.

The saltpool problem has been investigated by various authors with different success (e.g., Ackerer et al. [3], Thiele [512], Oswald and Kinzelbach [405, 406], Johannsen et al. [289], Diersch [135], Diersch and Kolditz [138], Häfner and Stüben [221], Häfner and Boy [220], Mazzia and Putti [359]). The numerical modeling is complicated due to the extremely small dispersivities and a large density contrast particularly for the high density case with 10 % mass fraction of salt. Good agreements with the measurements have been achieved by Johannsen et al. [289]. To fit both experiments, however, they had to adjust some parameters within accepted bounds given in parentheses: permeability (20 %), porosity (4 %) and transverse dispersivity (50 %). They studied mesh convergence by using a hierarchy of regular meshes consisting of hexahedral elements, up to mesh level $\ell = 8$, where the total number of elements is $N_{\mathrm{E}} = 8^{\ell}$. It was shown that extremely fine meshes (up to about 17 million nodal points with an element length of 0.78125 mm!) are required to model the high density case with sufficient accuracy. In addition, salinity-dependent viscosity effects had to be taken into account.

For the present FEFLOW computations meshes with only moderate resolutions are employed. We use both a structured mesh of hexahedral elements with a mesh level of $\ell = 6$ (mesh A consisting of 274,625 nodes) and an unstructured mesh of pentahedral elements for only the symmetric half, which is partially refined at the outlet (mesh B consisting of 140,010 nodes). For the computations the GFEM without any upwind and the AB/TR adaptive time stepping combined with a one-step Newton method were applied, thus ensuring that the numerical results will be second-order accurate, both in time and space. The model parameters are summarized in Table 11.10. Unspecified BC's represent no-flow boundaries, at which natural BC's are imposed.

The results for both the low (1) and high density cases (2) are presented in Fig. 11.39. It reveals the role of density effects in mixing and in dilution of saltwater, which is mainly controlled by the hydrodynamic dispersion process. In the high density case, the transition zone between saline and freshwater is significantly widened, forming a 'diffusive upcone' below the outlet at very low concentrations. This mixing process is considerably influenced by the advective and dispersive forces acting locally on the saltwater-freshwater interface, which is initially very narrow.

The simulation of the low density case agrees well with the measurements. However, differences in the long-term behavior remain (Fig. 11.39c, left). A previous solution based on the local smoothing technique of the velocity approximation completely failed for the high density case, as the saltwater mixing concentration at the outlet is significantly overestimated [135, 406]. Using FKA for the consis-

Table 11.10 Simulation parameters of the saltpool problem

		Value		
Quantity	Symbol	Low density	High density	Unit
Cell measure (width; depth; height)[a]	$B; D; H$	0.2; 0.2; 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
Isotropic hydraulic conductivity	K	$9.773 \cdot 10^{-3}$		$\mathrm{m\,s^{-1}}$
Porosity	ε	0.372		1
Specific storage coefficient	S_o	0		$\mathrm{m^{-1}}$
Molecular diffusion coefficient	D	$1 \cdot 10^{-9}$		$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	$1.2 \cdot 10^{-3}$		m
Transverse dispersivity	β_T	$1.2 \cdot 10^{-4}$		m
Reference salinity (freshwater)	C_0	0		$\mathrm{kg\,m^{-3}}$
Maximum salinity	C_s	10	100	$\mathrm{kg\,m^{-3}}$
Specific solutal expansion coefficient	α	$7.6 \cdot 10^{-3}$	$7.35 \cdot 10^{-2}$	1
Inflow/outflow rate	Q	$1.89 \cdot 10^{-6}$	$1.83 \cdot 10^{-6}$	$\mathrm{m^3\,s^{-1}}$
Variable liquid viscosity[b]	$f_\mu = \frac{\mu_0}{\mu}$	$= \frac{1}{1+1.85\omega-4.1\omega^2+44.5\omega^3}$		1
Flow BC				
Dirichlet-type BC at inlet $(x = 0; y = 0; z = H)$	h_D	0		m
Well-type SPC at outlet $(x = B; y = D; z = H)$	$Q_w = Q$	0.163296	0.158112	$\mathrm{m^3\,d^{-1}}$
Mass IC and BC's				
Initial condition (IC) of salinity	C_0	$\begin{cases} C_s & \text{for } z \leq H_2 \\ 0 & \text{for } z > H_2 \end{cases}$		$\mathrm{kg\,m^{-3}}$
Dirichlet-type BC at inlet $(x = 0; y = 0; z = H)$	$C_D = C_0$	0		$\mathrm{kg\,m^{-3}}$
FEM				
3D meshes of different resolutions, GFEM, AB/TR, OB approximation				
Initial time step size[c]	Δt_0	10^{-8}		d
RMS error tolerance (AB/TR)	ϵ	$5 \cdot 10^{-5}$		1
Simulation time period	t_{end}	160		min

[a] Measures defined in Fig. 11.38

[b] Using relationship (3.213) as function of mass fraction $\omega = \frac{C}{\rho_0}$, where $\rho_0 = \rho(C_0) = 10^3\,\mathrm{kg\,m^{-3}}$

[c] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$ and maximum time step size $\Delta t_{\max} = 2 \cdot 10^{-4}\,\mathrm{d}$

tent velocity approximation, the computation of the breakthrough curves is in a reasonable agreement with the experiment (Fig. 11.39c, right). This emphasizes the importance of a consistent velocity approximation for high density situations, which has proven to be a fundamental requirement for the successful solution of the saltpool problem at large density contrasts. Small inconsistencies in the velocity field would have dramatic consequences on the computational results. The data fit can be improved by re-adjusting parameters, in particular the transverse dispersivity β_T, porosity ε and conductivity K as shown by Johannsen et al. [289]. They have

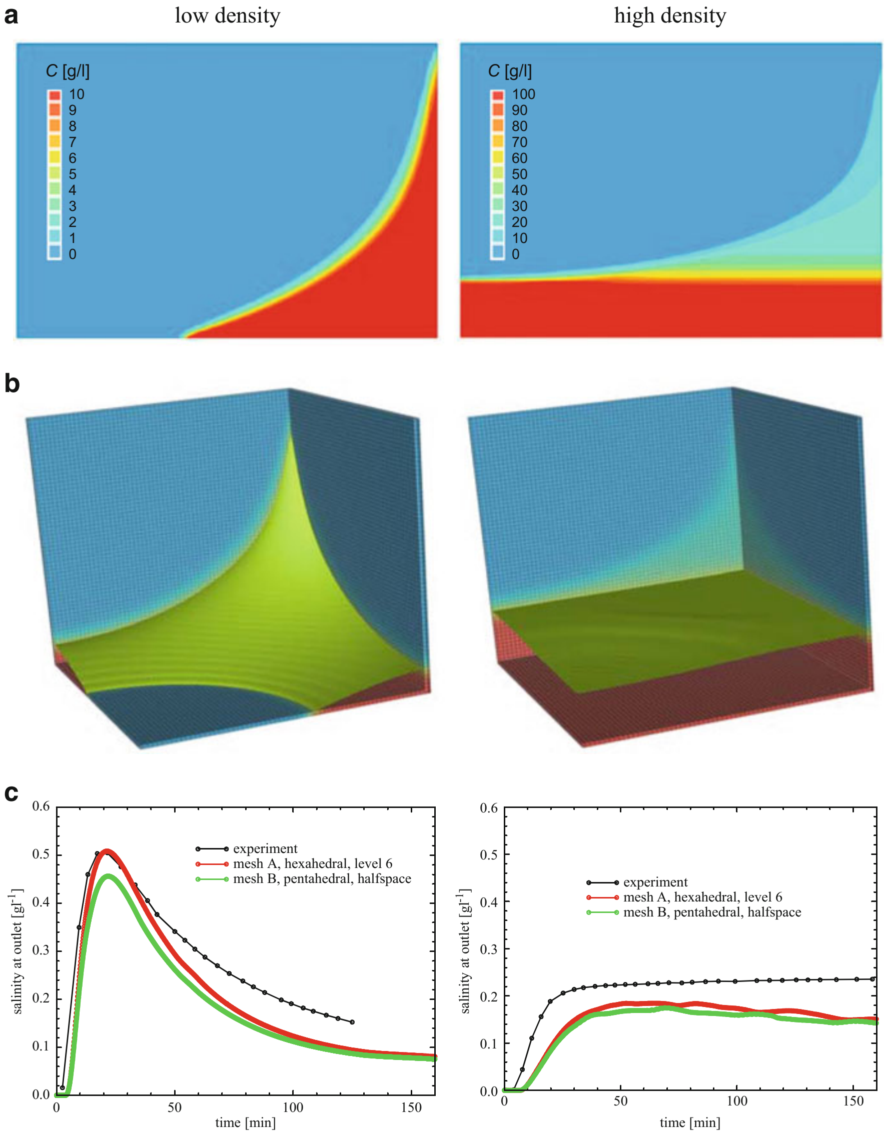


Fig. 11.39 FEFLOW results of the saltpool problem for the low (*left*) and high density cases (*right*): (**a**) cross-sectional salinity distribution, (**b**) 50 % salinity surface at $t = 160$ min, (**c**) salinity breakthrough curves at the outlet obtained for meshes A and B, *black circles* correspond to experimental results (Modified from [138])

also shown in their mesh convergence study that a mesh level of $\ell = 6$ represents a minimum spatial resolution required for an accurate simulation of the high density case.

11.11.8 Pringle et al.'s Double-Diffusive Finger Convection Problem

Double-diffusive finger convection (DDFC) phenomena represent supercritical convection regimes, which are characterized by long fingering patterns of rising and falling fluid (cf. Sect. 11.5.3.2). It has been recognized that subsurface environments (porous media and fractures) are favorable to DDFC. It can be important at deep circulation in marine and terrestrial alluvial basins, for interaction of groundwater and surface water and in transport of dissolved solutes from solid waste landfills.

DDFC processes were studied by Cooper et al. [98] and Pringle et al. [431] via Hele-Shaw experiments using a light transmission technique that provides high-resolution concentration fields. From a near perturbation-free initial layering of a lighter sucrose solution over a dense salt solution, upward and downward moving fingers quickly form at the interface between the two solutions. Particularly, the recent experimental data obtained by Pringle et al. [431] provide a suited baseline for use in the development and evaluation of numerical models.

Numerical models must play an increasing role in a better understanding of DDFC phenomena in porous media. As already argued by Cooper et al. [98] a limitation in finger growth due to large-scale circulation controlled by inertial forces as observed in ordinary fluids (nonporous media) does not seem to exist in porous systems characterized by low Reynolds numbers. An intriguing possibility is that the merging and subsequent formation of conduits along which fingers travel could be repeated at larger and larger scales. Cooper et al. [98] concluded that larger and greater-spaced conduits for mass transport may naturally evolve in porous media, leading to growth bounded on a much larger scale than has been observed in any laboratory experiments.

The Hele-Shaw experiments collected by Pringle et al. [431] were successfully simulated by Hughes et al. [271] using a modified version of the SUTRA code [270, 550] that combines GFEM and integrated FDM. The dataset of Pringle et al. [431] is well-suited for code verification of DDFC numerical models because, unlike most previous experimental Hele-Shaw datasets, it is of sufficient spatial and temporal resolution to allow accurate comparisons of simulated and measured convective fingering. In addition, computational high-resolution results obtained by different numerical approaches and full-field images from the experimental dataset allow qualitative comparison of the evolving flow field and quantitative comparison of mass transfer rates.

Pringle et al. [431] used a Hele-Shaw cell to study the temporal and spatial distribution of DDFC phenomena of two solutes initially in a density-stable configuration with a mean interface thickness of about $1 \cdot 10^{-3}$ m. The Hele-Shaw cell was filled with a sucrose solution over a denser sodium chloride solution (NaCl). The 2D domain measures 0.2541 m (cell length L) by 0.1625 m (cell height H), see Fig. 11.40. The Hele-Shaw cell is inclined at an angle of 25° relative to horizontal. To visualize sodium chloride concentrations and quantify convective motion, a dye tracer with a low concentration was mixed with the

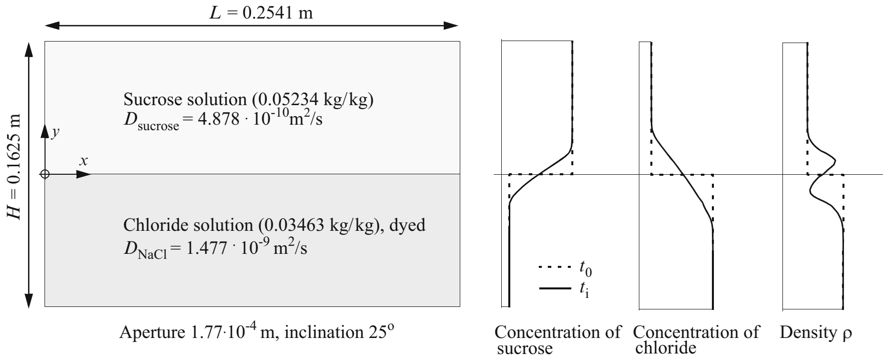

L = 0.2541 m
H = 0.1625 m
y
x
Sucrose solution (0.05234 kg/kg)
$D_{\text{sucrose}} = 4.878 \cdot 10^{-10}$ m²/s
Chloride solution (0.03463 kg/kg), dyed
$D_{\text{NaCl}} = 1.477 \cdot 10^{-9}$ m²/s
Aperture 1.77·10⁻⁴ m, inclination 25°
t_0
t_i
Concentration of sucrose
Concentration of chloride
Density ρ

Fig. 11.40 Hele-Shaw study experiment by Pringle et al. [431]

sodium chloride. The dye had a negligible effect on fluid density. Accordingly, three species ($N^\star = 3$) have to be considered: sucrose ($k = s$), sodium chloride ($k = c$) and dye ($k = d$). The used parameters are summarized in Table 11.11. Unspecified BC's represent no-flow boundaries, at which natural BC's are imposed. Note that NaCl concentrations are not mapped perfectly by the dye tracer because the diffusivity of sodium chloride is approximately 2.5 times greater than the diffusivity of the dye (Table 11.11). Because the motion is convective through most of the experiment, Pringle et al. [431] suggested the diffusivity differences had little impact on the mapping of sodium chloride concentrations over the length of time of the experiment. Note further, the liquid viscosity is approximated as a linear function of solute concentration

$$f_\mu = \frac{\mu_0}{\mu(C_k)} \approx \frac{\mu_0}{\mu_0 + \sum_k^{N^\star} \vartheta_k (C_k - C_{k0})} \tag{11.96}$$

where the viscosity change coefficients ϑ_k are given in Table 11.11 for each species.

To maintain the full physical equivalence to the experimental and numerical studies done by Pringle et al. [431] and Hughes et al. [271] the most important physical quantities characterizing the DDFC problem are the Turner number Tu given by 1.22, the Lewis number Le according to 0.3303 and one Rayleigh number given for sodium chloride as $\text{Ra}_c = 26{,}460$. The remaining quantities can be directly derived from these characteristic numbers. Note that by using the dimensionless density expansion coefficients α_k in the fluid buoyancy χ of (11.2), the density expansion becomes independent of the real values of species concentrations and the maximum concentrations C_{ks} can be arbitrarily chosen. In agreement to the physical experiment C_{ks} and C_{k0} are chosen as listed in Table 11.11.

It has been shown in stability analysis [389] and Hele-Shaw experiments [97] as the Turner number Tu decreases from the stability boundary at Le^{-1}, the system transitions from being diffusion-dominated to convection-dominated. The

Table 11.11 Parameters of Pringle et al.'s DDFC problem

Quantity	Symbol	Value	Unit
Cell measure (length; height)[a]	L; H	0.2541; 0.1625	m
Cell plate spacing	b	$1.77 \cdot 10^{-4}$	m
Cell angle to the horizontal	θ	25	°
Effective gravity	$g_y = g \sin\theta$	4.14	$\mathrm{m\,s^{-2}}$
Reference concentrations	C_{k0} $(k = s, c, d)$	0	$\mathrm{kg\,m^{-3}}$
Maximum sucrose concentration	C_{ss}	52.235	$\mathrm{kg\,m^{-3}}$
Maximum chloride concentration	C_{cs}	34.561	$\mathrm{kg\,m^{-3}}$
Maximum dye concentration	C_{ds}	0.2495	$\mathrm{kg\,m^{-3}}$
Reference liquid density	$\rho_0 = \rho(C_{s0}, C_{c0}, C_{d0})$	998	$\mathrm{kg\,m^{-3}}$
Sucrose expansion coefficient	α_s	0.0182787	1
Chloride expansion coefficient	α_c	0.022302	1
Dye expansion coefficient	α_d	0	1
Dynamic viscosity of liquid	μ_0	$1 \cdot 10^{-3}$	$\mathrm{kg\,m^{-1}\,s^{-1}}$
Cell intrinsic permeability	$k = b^2/12$	$2.61 \cdot 10^{-9}$	$\mathrm{m^2}$
Isotropic hydraulic conductivity	$K = \frac{k\rho_0 g_y}{\mu_0}$	$1.07838 \cdot 10^{-2}$	$\mathrm{m\,s^{-1}}$
Porosity	ε	1	1
Specific storage coefficient	S_o	0	$\mathrm{m^{-1}}$
Sucrose diffusion coefficient	D_s	$4.878 \cdot 10^{-10}$	$\mathrm{m^2\,s^{-1}}$
Chloride diffusion coefficient	D_c	$1.477 \cdot 10^{-9}$	$\mathrm{m^2\,s^{-1}}$
Dye diffusion coefficient	D_d	$5.670 \cdot 10^{-10}$	$\mathrm{m^2\,s^{-1}}$
Dispersivities	β_L; β_T	0; 0	m
Viscosity change to sucrose[b]	ϑ_s	$2.75 \cdot 10^{-3}$	$\mathrm{m^2\,s^{-1}}$
Viscosity change to chloride[b]	ϑ_c	$1.59 \cdot 10^{-3}$	$\mathrm{m^2\,s^{-1}}$
Viscosity change to dye[b]	ϑ_d	0	$\mathrm{m^2\,s^{-1}}$
Sucrose Rayleigh number	$\mathrm{Ra}_s = \frac{\alpha_s K H}{\varepsilon D_s}$	65,664	1
Chloride Rayleigh number	$\mathrm{Ra}_c = \frac{\alpha_c K H}{\varepsilon D_c}$	26,460	1
Lewis number	$\mathrm{Le} = \frac{D_s}{D_c}$	0.3303	1
Turner number	$\mathrm{Tu} = \frac{\alpha_c}{\alpha_s}$	1.22	1
Flow BC			
Dirichlet-type BC at central point	$h(x, y) = h_D(\frac{L}{2}, 0)$	0	m
Species IC's			
Sucrose IC[c]	C_{s0}	$\begin{cases} C_{ss} & \text{for } y \geq 0 \\ 0 & \text{for } y < 0 \end{cases}$	$\mathrm{kg\,m^{-3}}$
Chloride IC[c]	C_{c0}	$\begin{cases} C_{cs} & \text{for } y \leq 0 \\ 0 & \text{for } y > 0 \end{cases}$	$\mathrm{kg\,m^{-3}}$
Dye IC	C_{d0}	$\begin{cases} C_{ds} & \text{for } y \leq 0 \\ 0 & \text{for } y > 0 \end{cases}$	$\mathrm{kg\,m^{-3}}$
FEM			
Uniform 2D meshes of different resolutions, FE/BE, GFEM, OB and EOB approximation			
Initial time step size	Δt_0	10^{-8}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Simulation time period[d]	$\hat{t}_{\text{end}}$	$3.17 \cdot 10^{-3}$	1

[a] Measures defined in Fig. 11.40

[b] Using viscosity relation function $f_\mu = \frac{\mu_0}{\mu} = \frac{\mu_0}{\mu_0 + \vartheta_s(C_s - C_{s0}) + \vartheta_c(C_c - C_{c0}) + \vartheta_d(C_d - C_{d0})}$

[c] Concentrations at the interface points $(0 \leq x \leq L, y \equiv 0)$ are disturbed according to (11.99)

[d] Dimensionless time: $\hat{t} = \frac{D_c}{H^2} t$

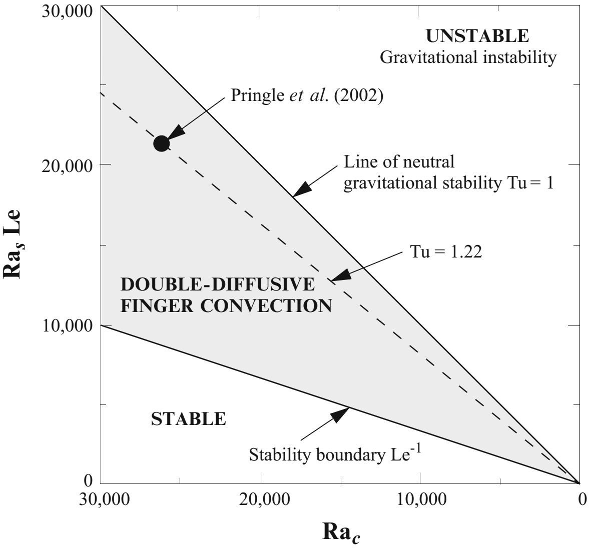


Fig. 11.41 Stability and instability domains in the Rayleigh parameter space with the location of the present DDFC problem at Tu = 1.22 (Modified from [271])

corresponding stability and instability domains in the Rayleigh parameter space are shown in Fig. 11.41. The current situation with a Turner number of 1.22 is clearly located in the DDFC domain with increasing mass fluxes and finger velocities. DDFC exists in the range $1 < \mathrm{Tu} < \mathrm{Le}^{-1}$. For $\mathrm{Tu} < 1$ there is no more an initially density-stable stratification of the solutes and the system becomes gravitationally instable.

In the 16-h Hele-Shaw experiments done by Pringle et al. [431] a total of 300 images of the evolving concentration field was collected. A sequence of dye concentrations from the experiment is shown in Fig. 11.42. Time is presented as dimensionless $\hat{t} = \frac{D_c}{H^2} t$. The measured time stages (t and $\hat{t}$) are listed in Table 11.12.

As seen in a sequence of concentration fields in Fig. 11.42 there are interesting features in the behavior of the DDFC system. Due to the initially perturbed solution interface an array of distinct fingers rapidly grows in unison at the early time stage (Fig. 11.42a, b). These fingers begin to interact with one another causing a re-organization of the initial uniform finger structure (Fig. 11.42c, d). A typical feature at this stage is a large number of very small fingers with a wide variation in vertical extent. As convection proceeds, small-scale fingers continuously emerge from the region of the initial solution interface referred to as the *finger generation zone* by Cooper et al. [98]. These newly generated fingers add to the structural intricacy of the field by growing, and in many cases, merging with, and convecting up through the stems of early formed neighbors. The generation of new finger pairs as the tips of some upward and downward growing fingers can also be observed (Fig. 11.42c–f). At $\hat{t} = 4.23 \cdot 10^{-4}$ (Fig. 11.42e), the fastest growing fingers reach the top and bottom boundaries of the cell and begin to spread laterally forming more dense (at the bottom) and less dense (at the top) 'clouds' of fluid (Fig. 11.42f–h). Within the finger generation zone, far from the boundaries, new fingers continue

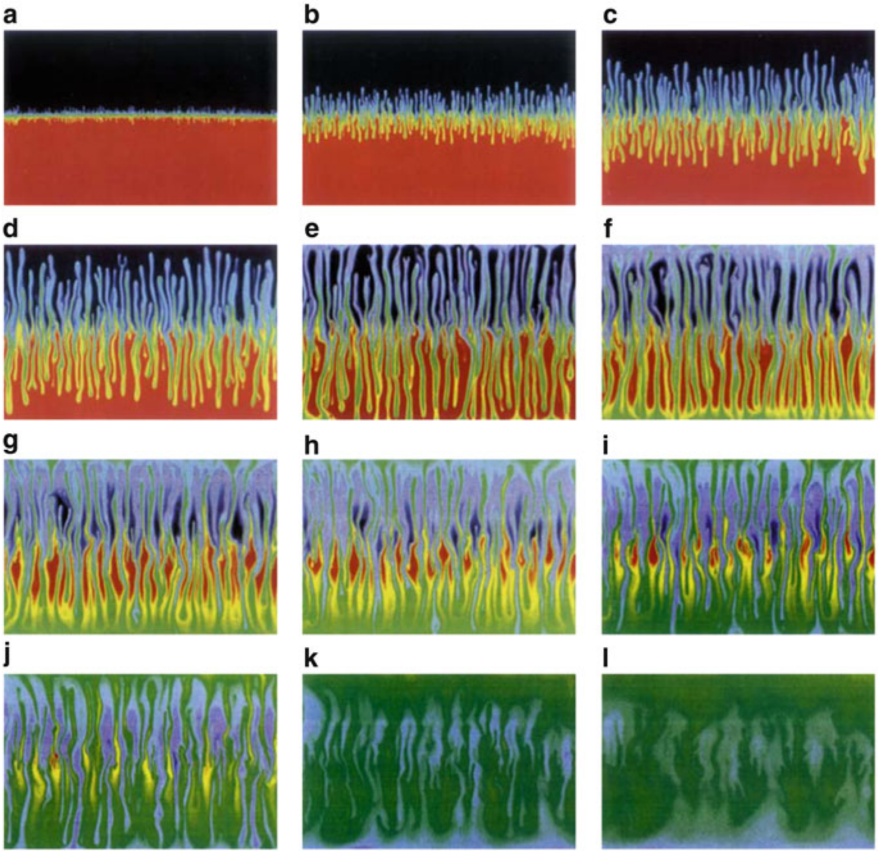


Fig. 11.42 Hele-Shaw observation results from Pringle et al. [431] for the dye component at (**a**) $\hat{t} = 4.03 \cdot 10^{-5}$, (**b**) $\hat{t} = 1.31 \cdot 10^{-4}$, (**c**) $\hat{t} = 2.21 \cdot 10^{-4}$, (**d**) $\hat{t} = 3.22 \cdot 10^{-4}$, (**e**) $\hat{t} = 4.23 \cdot 10^{-4}$, (**f**) $\hat{t} = 5.24 \cdot 10^{-4}$, (**g**) $\hat{t} = 6.04 \cdot 10^{-4}$, (**h**) $\hat{t} = 7.25 \cdot 10^{-4}$, (**i**) $\hat{t} = 7.85 \cdot 10^{-4}$, (**j**) $\hat{t} = 1.03 \cdot 10^{-3}$, (**k**) $\hat{t} = 1.77 \cdot 10^{-3}$ and (**l**) $\hat{t} = 3.17 \cdot 10^{-3}$. ($\hat{t} = \frac{D_c}{H^2} t$ dimensionless time). Color sequence *black-blue-green-yellow-orange-red* depicts normalized dye concentration from 0 to 1

to form from isolated pockets of nearly pristine solution located about the initial solution interface (Fig. 11.42g–j). Finally, at late time, the finger structure becomes 'tree-like' with a branching pattern that has greater lateral travel than at early time. This final convective structure remains long after motion has stopped, diffusion now acting to slowly uniformize the field (Fig. 11.42l).

To solve successfully the DDFC problem a sufficiently fine spatial discretization is fundamental. Because most transfer in a DDFC system is a result of convection, small finger dimensions may evolve. Damping effects by artificial numerical dispersion should be hold down on a lowest level to resolve accurately the minimum finger dimension occurring in a DDFC simulation. A further important point in DDFC computations refers to arising numerical perturbations which can affect the evolution of DDFC [138]. It is to be expected that uniform and aligned structured

Table 11.12 Measured time stages

	Stages	$\hat{t}$ (-)	t (s)	t (h)
(a)	*Early* stage	$4.03 \cdot 10^{-5}$	720.5	0.20
(b)	*Mature* stage: vertical growth of fingers	$1.31 \cdot 10^{-4}$	2,342.1	0.65
(c)		$2.21 \cdot 10^{-4}$	3,951.1	1.10
(d)		$3.22 \cdot 10^{-4}$	5,756.8	1.60
(e)	Fingers reach top and bottom boundaries	$4.23 \cdot 10^{-4}$	7,562.5	2.10
(f)		$5.24 \cdot 10^{-4}$	9,368.2	2.60
(g)	*Roundown* stage	$6.04 \cdot 10^{-4}$	10,798.5	3.00
(h)		$7.25 \cdot 10^{-4}$	12,961.8	3.60
(i)		$7.85 \cdot 10^{-4}$	14,034.5	3.90
(j)		$1.03 \cdot 10^{-3}$	18,414.7	5.12
(k)		$1.77 \cdot 10^{-3}$	31,644.6	8.79
(l)		$3.17 \cdot 10^{-3}$	56,674.2	15.7

Table 11.13 Meshes according to refinement levels ℓ

Level ℓ	Used FEFLOW mesh	N_E	N_P	Spatial increment (mm)
0	–	41,984 (256 × 164)	42,405 (257 × 165)	0.992
1	–	67,936 (512 × 328)	168,777 (513 × 329)	0.496
2	**mesh A**	671,744 (1,024 × 656)	673,425 (1,025 × 657)	0.248
3	**mesh B**	2,686,976 (2,048 × 1,312)	2,690,337 (2,049 × 1,313)	0.124
4	–	10,747,904 (4,096 × 2,624)	10,754,625 (4,097 × 2,625)	0.062

meshes with square elements can minimize uncontrollable numerical perturbations during the simulation.

For the present computations quadrilateral meshes with different resolution are applied. It can be recognized as a stepwise global refinement of meshing: starting with a 2D discretization each quadrilateral is subdivided into four equally sized quadrilaterals. The number of linear quadrilateral elements N_E and number of nodes N_P then increase according to the refinement level $\ell = 0, 1, 2, \ldots$:

$$\begin{aligned} N_\mathrm{E} &= 41 \cdot 2^{(10+2\ell)} \\ N_\mathrm{P} &= N_\mathrm{E} + 105 \cdot 2^{(2+\ell)} + 1 \end{aligned} \tag{11.97}$$

Table 11.13 summarizes the mesh properties up to level 4.

Hughes et al. [271] simulated meshes at levels ℓ of 0, 1 and 2. Their computations with the finest mesh at $\ell = 2$ agreed rather well with Pringle et al.'s Hele-Shaw

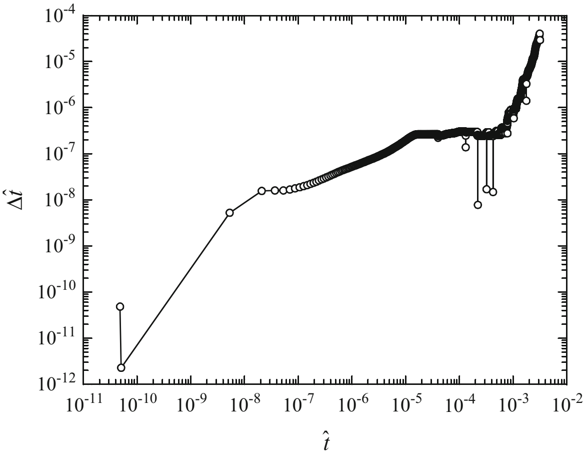


Fig. 11.43 Time stepping history for mesh B

experiments. They found that the coarser discretizations with $\ell = 0$ and $\ell = 1$ are inappropriate to model the finger development with a reasonable accuracy. However, even their finest discretization at $\ell = 2$ with a spatial increment of 0.248 mm is still larger than the pixel size with 0.154 mm of the Hele-Shaw experiment by a factor of 1.6. More refined meshes could not be simulated by Hughes et al. [271] due to their computational limitations.

In the present FEFLOW simulations we recompute the DDFC problem in using Hughes et al.'s finest 671,744−element mesh at $\ell = 2$. Additionally, FEFLOW simulations are performed on a further refined mesh having the refinement level $\ell = 3$. In the following FEFLOW simulations we denote these meshes as mesh A consisting of 1,024 × 656 linear quadrilateral elements (673,424 nodes) and mesh B consisting of 2,048 × 1,312 linear quadrilateral elements (2,690,337 nodes), see Table 11.13. Mesh A is comparable to the finest spatial discretization used by Hughes et al. [271]. Note that the high-resolution mesh B is more refined than the length scales of in the Hele-Shaw experiment. The spatial increment in mesh B with 0.124 mm is smaller than the pixel size of the Hele-Shaw experiment with 0.154 mm.

It is important to note that mesh B requires 64-bit execution. In the present study we prefer the FE/BE predictor-corrector time stepping strategy and parallel computations. While the flow equations are solved by using AMG equation solver (Sect. 8.17.2.7), the species transport equations are solved by using BiCGSTAB equation solver (Sect. 8.17.2.6) with ILU preconditioning. Both solvers are applied with a reduced stop criteria of 10^{-12} to terminate iterations in solving the sparse finite-element matrix equation systems.

All external boundary faces represent no-flux conditions both for fluid flow and for species mass transport. This is automatically satisfied by natural (zero-value) Neumann-type BC's and no specifications are required. But, there is one exception. Because the specific storage coefficient S_o is zero in the flow equation there is no

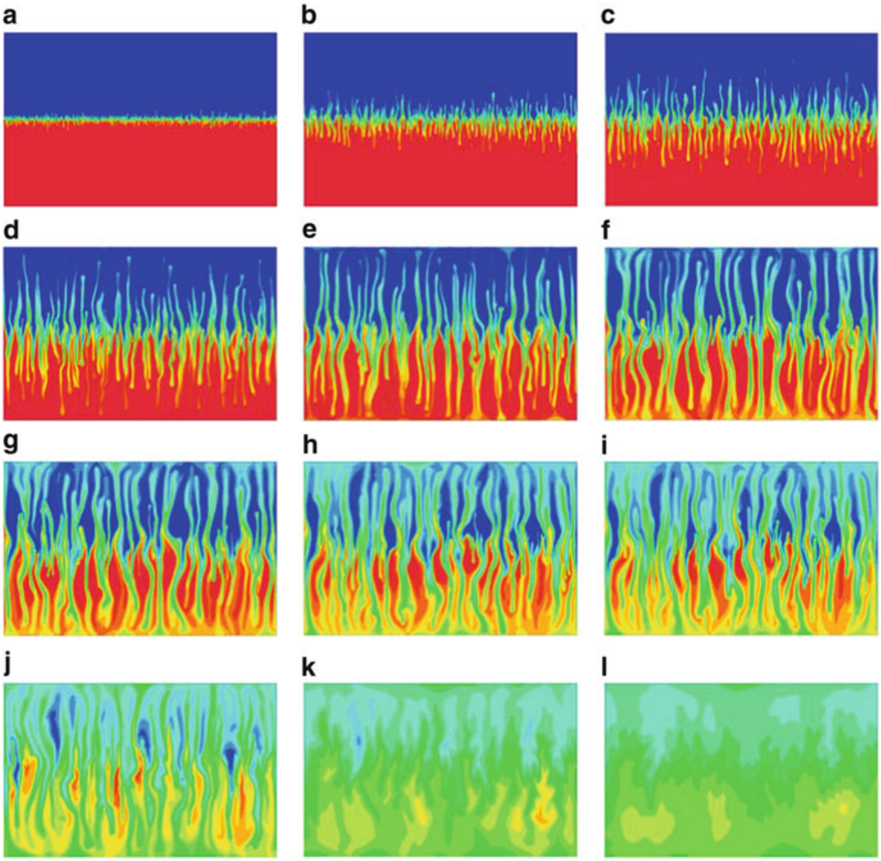


Fig. 11.44 FEFLOW results simulated with mesh A for the dye component at (**a**) $\hat{t} = 4.03 \cdot 10^{-5}$, (**b**) $\hat{t} = 1.31 \cdot 10^{-4}$, (**c**) $\hat{t} = 2.21 \cdot 10^{-4}$, (**d**) $\hat{t} = 3.22 \cdot 10^{-4}$, (**e**) $\hat{t} = 4.23 \cdot 10^{-4}$, (**f**) $\hat{t} = 5.24 \cdot 10^{-4}$, (**g**) $\hat{t} = 6.04 \cdot 10^{-4}$, (**h**) $\hat{t} = 7.25 \cdot 10^{-4}$, (**i**) $\hat{t} = 7.85 \cdot 10^{-4}$, (**j**) $\hat{t} = 1.03 \cdot 10^{-3}$, (**k**) $\hat{t} = 1.77 \cdot 10^{-3}$ and (**l**) $\hat{t} = 3.17 \cdot 10^{-3}$. ($\hat{t} = \frac{D_c}{H^2} t$ dimensionless time). Color sequence *blue-green-yellow-orange-red* depicts normalized dye concentration from 0 to 1

more a regular time-derivative term and the flow equations should be linked to a Dirichlet-type BC to stabilize the numerical solution. It is sufficient to specify at least one node with an arbitrary head value h. While Hughes et al. [271] specified both the upper left and upper right corner nodes with values for pressure and species concentrations, in our simulations only the node at the center of the mesh is specified with a hydraulic head h of 0.0, no extra BC's are introduced for the species concentrations.

At initial time t_0 the three species $k = s, c, d$ are distributed as follows within the 2D domain ($0 \leq x \leq L, -\frac{H}{2} \leq y \leq \frac{H}{2}$) in a layered configuration, where the solute interface is located at $y = 0$ (see Fig. 11.40):

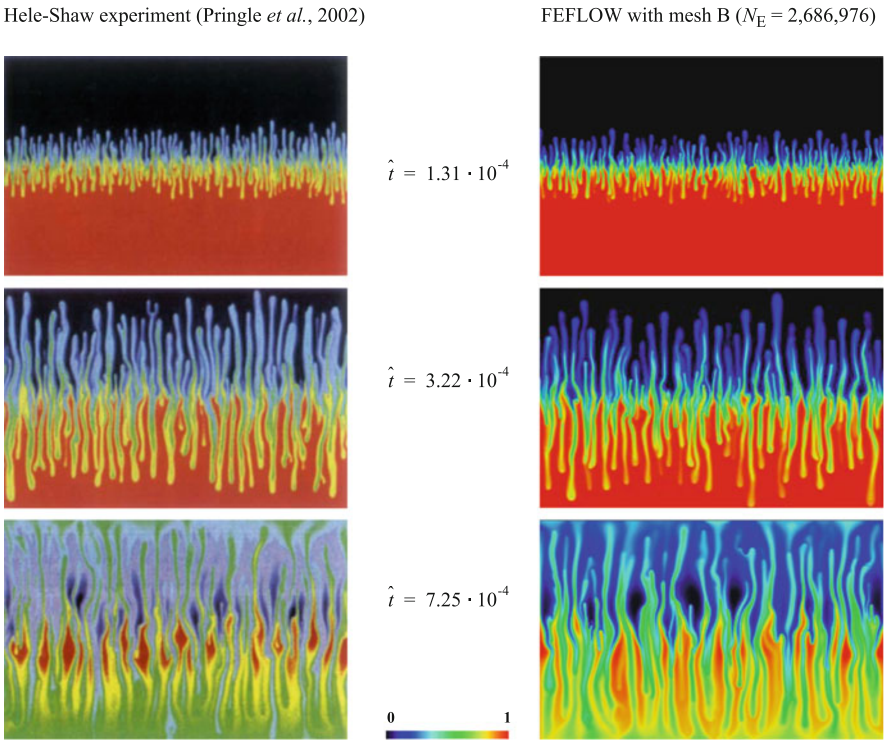


Fig. 11.45 Comparison of Hele-Shaw experiments from Pringle et al. [431] to FEFLOW results simulated with mesh B for the dye component at different dimensionless times $\hat{t} = \frac{D_c}{H^2} t$. Case of variable fluid viscosity $f_\mu = \mu_0/\mu(C_k)$. Color sequence *black-blue-green-yellow-orange-red* depicts normalized dye concentration from 0 to 1

$$
\begin{aligned}
C_s(\boldsymbol{x}, t_0) &= \begin{cases} C_{ss}(0 \le x \le L, 0 \le y \le \frac{H}{2}) \\ C_{s0}(0 \le x \le L, -\frac{H}{2} \le y < 0) \end{cases} \\
C_c(\boldsymbol{x}, t_0) &= \begin{cases} C_{c0}(0 \le x \le L, 0 < y \le \frac{H}{2}) \\ C_{cs}(0 \le x \le L, -\frac{H}{2} \le y \le 0) \end{cases} \\
C_d(\boldsymbol{x}, t_0) &= \begin{cases} C_{d0}(0 \le x \le L, 0 < y \le \frac{H}{2}) \\ C_{ds}(0 \le x \le L, -\frac{H}{2} \le y \le 0) \end{cases}
\end{aligned}
\qquad (11.98)
$$

The present finger convection problem is very sensitive with respect to perturbations. Pringle et al. [431] expended significant effort in minimizing initial perturbations for the Hele-Shaw experiment. Although the thickness of the solute interface was small (about 1 mm), perturbations at the start of the experiment could not be avoided. They were seeds for initial finger developments. For the numerical simulation a control of such type of initial seeds for finger developments is needed. This should be mimicked by the following random procedure as proposed by Hughes et al. [271].

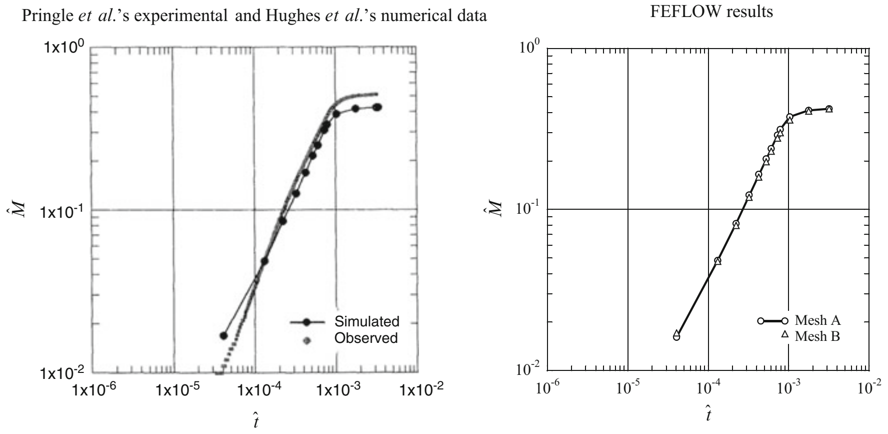


Fig. 11.46 Normalized mass transfer across the center line $\hat{M}$. Comparison of observed data taken by Pringle et al. [431] and numerical results by Hughes et al. [271] (*left*) with FEFLOW results computed for meshes A and B (*right*)

Random noise with a mean of zero and maximum amplitude of 0.5 % of maximum initial concentrations C_{ks} is applied to both sucrose and sodium chloride at the initial solution interface. Dye concentrations at the interface are not perturbed. To develop initial perturbations for sucrose and sodium chloride their nodal concentrations at nodes sharing the solute interface at $y = 0$ are modified as follows:

```
Do for all interface nodes i {
   RN1 = random number between 0 and 1
   If (RN1 < 0.5) {
      $C_s(x_i, y_i = 0, t_0) = 0.01 \cdot \mathrm{RN1} \cdot C_{ss}$
   }
   Else {
      $C_s(x_i, y_i = 0, t_0) = C_{ss}$
   }

   RN2 = random number between 0 and 1
   If (RN2 < 0.5) {
      $C_c(x_i, y_i = 0, t_0) = 0.01 \cdot \mathrm{RN2} \cdot C_{cs}$
   }                                                        (11.99)
   Else {
      $C_c(x_i, y_i = 0, t_0) = C_{cs}$
   }
}
```

where x_i and y_i correspond to the $x-$ and $y-$coordinates of node i.

Table 11.14 Measured vs. simulated $\hat{M}$

		$\hat{M}$	
	$\hat{t}$	Hele-Shaw experiment Pringle et al. [431]	FEFLOW mesh B
(a)	$4.03 \cdot 10^{-5}$	0.01	0.02
(b)	$1.31 \cdot 10^{-4}$	0.05	0.05
(c)	$2.21 \cdot 10^{-4}$	0.10	0.08
(d)	$3.22 \cdot 10^{-4}$	0.15	0.12
(e)	$4.23 \cdot 10^{-4}$	0.20	0.16
(f)	$5.24 \cdot 10^{-4}$	0.25	0.20
(g)	$6.04 \cdot 10^{-4}$	0.30	0.23
(h)	$7.25 \cdot 10^{-4}$	0.35	0.28
(i)	$7.85 \cdot 10^{-4}$	0.40	0.30
(j)	$1.03 \cdot 10^{-3}$	0.45	0.36
(k)	$1.77 \cdot 10^{-3}$	0.50	0.41
(l)	$3.17 \cdot 10^{-3}$	0.51	0.42

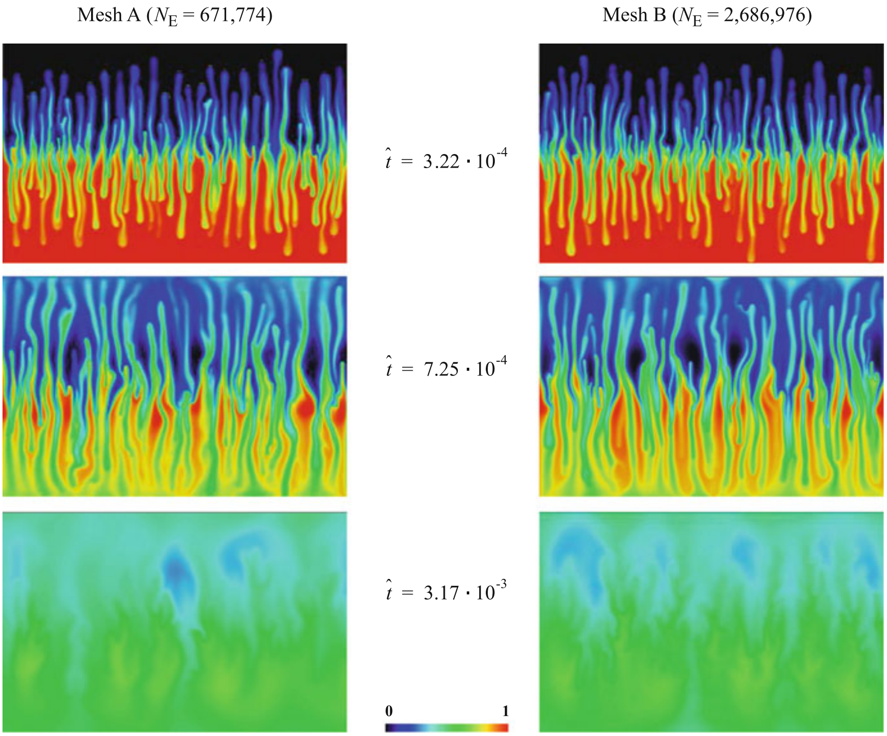


Fig. 11.47 FEFLOW results simulated for the dye component at different dimensionless times $\hat{t} = \frac{D_c}{H^2} t$. Comparison between mesh A (*left*) and mesh B (*right*) for the case of variable fluid viscosity $f_\mu = \mu_0 / \mu(C_k)$. Color sequence *black-blue-green-yellow-orange-red* depicts normalized dye concentration from 0 to 1

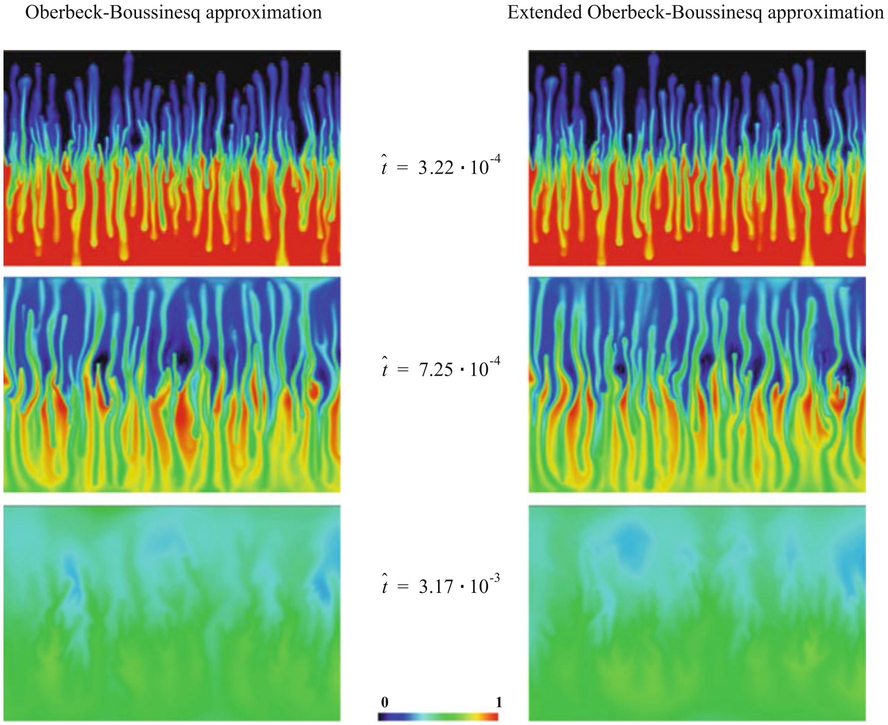


Fig. 11.48 FEFLOW results simulated with mesh A for the dye component at different dimensionless times $\hat{t} = \frac{D_c}{H^2} t$. Comparison of OB approximation $Q_{\mathrm{EOB}} \equiv 0$ (*left*) to the EOB approximation $Q_{\mathrm{EOB}} \neq 0$ (*right*) for the case of constant fluid viscosity $f_\mu \equiv 1$. Color sequence *black-blue-green-yellow-orange-red* depicts normalized dye concentration from 0 to 1

The simulation of mesh A and B required 3,205 and 3,626 adaptive time steps, respectively. The time step history for the mesh B simulation is plotted in Fig. 11.43. Figure 11.44 shows the FEFLOW-simulated dye concentrations for mesh A at the same dimensionless times of Pringle et al. [431] (cf. Table 11.12). The results agree rather well with the computations obtained by Hughes et al. [271]. Qualitatively, the numerical results are similar to the experimental results as seen in Fig. 11.45 in comparison to the mesh B results. As already indicated by Hughes et al. [271] the experimental vertical finger evolution appears to be slightly ahead of the simulated fingers.

A more quantitative comparison can be done by using the vertical mass flux exemplified for the dye concentrations. A normalized mass transfer of dye upward across the centerline of the cell can be defined according to

$$\hat{M} = \frac{M}{M_0} \tag{11.100}$$

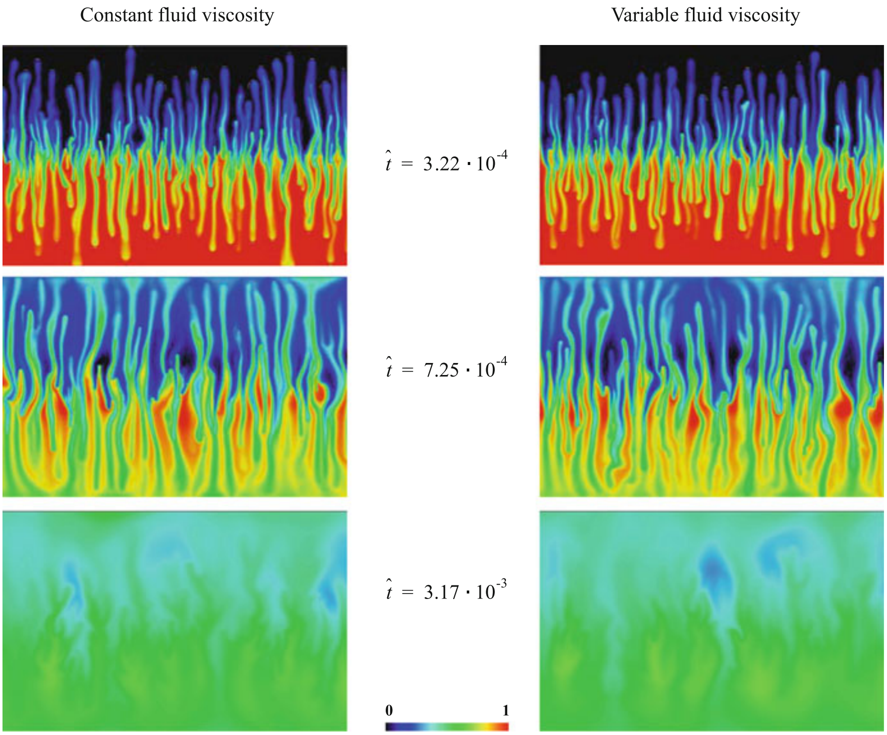


Fig. 11.49 FEFLOW results simulated with mesh A for the dye component at different dimensionless times $\hat{t} = \frac{D_c}{H^2} t$. Comparison of constant fluid viscosity $f_\mu \equiv 1$ (*left*) to the variable fluid viscosity case $f_\mu = \mu_0/\mu(C_k)$ (*right*). Color sequence *black-blue-green-yellow-orange-red* depicts normalized dye concentration from 0 to 1

where M is the dye mass above the centerline of the Hele-Shaw cell at time $\hat{t}$ and M_0 is the total dye mass in the cell. Numerical results compare reasonably well to observed values of $\hat{M}$ as depicted in Fig. 11.46. As also seen there FEFLOW's and Hughes et al.'s results agree very well. Their agreement with the Hele-Shaw experiment is acceptable until $\hat{t} = 1 \cdot 10^{-3}$. After $\hat{t} = 1 \cdot 10^{-3}$, the simulated mass transfer $\hat{M}$ is less than observed mass transfer. Larger percent errors at early times are an artifact of small $\hat{M}$ values and represent small absolute differences in mass transfer (e.g., 0.011 observed and 0.017 simulated) influenced by the initial perturbation at the interface nodes for the given spatial discretization. Note further that the simulated mass transfer $\hat{M}$ did not changed anymore if using a more refined mesh (cf. mesh A and mesh B results in Fig. 11.46). It indicates that the numerical accuracy with respect to the mass transfer is sufficiently achieved at a lower refinement level as given for mesh A. Table 11.14 compares the measured mass transfer $\hat{M}$ against the FEFLOW results obtained for mesh B.

A comparison of the finger evolution for the two meshes A and B is exhibited in Fig. 11.47. It reveals a slightly faster finger development for the more refined mesh

B compared to the coarser mesh A. While for mesh A at the front of the fingers small wiggles in the numerical solution could be observed at early times (indicated by white color spots in the fringed distributions of Fig. 11.47 left), the solution for mesh B is fully wiggle-free.

We also studied the influence of the OB approximation and the fluid viscosity on the simulation results. Noticeable but not significant differences exist in the simulated finger patterns when comparing the solutions with and without the OB approximation as seen in Fig. 11.48. More influence on the finger pattern results from the fluid viscosity effect. As evidenced in Fig. 11.49 a constant viscosity solution produces a slightly faster finger development as for the case with a variable (concentration-dependent) viscosity.

Chapter 12
Mass Transport in Porous Media with and Without Chemical Reactions

12.1 Introduction

In this chapter the computation of multispecies (including single-species) mass transport in porous media with chemical reaction in particular is examined. The complexity of those reactive transport processes arising in natural and engineered porous media requires some specific treatment due to their nonlinearity and the occurrence of multiple unknowns. In the preceding Chap. 5 the constitutive relations in form of reversible reaction and irreversible chemical kinetics have been developed. It ends up with a set of mass transport equations for each chemical species $k = 1, \ldots, N$ of an arbitrary number, nonlinearly coupled by the rate expressions of chemical reaction in form of degradation type, Arrhenius type, Monod type or freely editable kinetics. A given species k can be either *mobile* associated with a liquid (aqueous) phase l or *immobile* associated with a solid phase s, so that $N = N^l + N^s$. Chemicals in the liquid phase are subject to advection and dispersion, while in a solid phase there is no advection and dispersion. We solve the reactive multispecies mass transport processes in multi-dimensional porous media under variably saturated, variable-density and nonisothermal conditions. The focus of this chapter is on the treatment of the species mass transport PDE system, while for the flow computations we refer to Chap. 9 for saturated porous media, to Chap. 10 for variably saturated porous media and to Chap. 11 for density-coupled problems. Nonisothermal aspects are subject of Chaps. 11 and 13.

12.2 Basic Equations

12.2.1 3D, Vertical 2D and Axisymmetric Problems

The system of the basic PDE's for 3D and vertical 2D (including axisymmetric) multispecies mass transport in porous media has been developed in Sects. 3.10.5

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_12,

and 5.4 and summarized in Table 3.7. Due to the chemical reaction the equations can be nonlinearly coupled by the kinetic rate laws. The following general system of PDE's results for species $k = 1, \dots, N^l + N^s$ written for the divergence form of the mass transport equations

$$\begin{aligned}
\frac{\partial}{\partial t}(\varepsilon s \mathfrak{R}_k C_k) + \nabla \cdot (\boldsymbol{q} C_k) - \nabla \cdot (\boldsymbol{D}_k \cdot \nabla C_k) + \varepsilon s \vartheta_k \mathfrak{R}_k C_k &= \hat{R}_k + Q_{kw} + Q_k \\
&\quad \text{species } k \text{ of liquid phase } l \\
\frac{\partial}{\partial t}(\varepsilon_s C_k^s) + \varepsilon_s \vartheta_k C_k^s &= \hat{R}_k + Q_k \\
&\quad \text{species } k \text{ of solid phase } s
\end{aligned} \tag{12.1}$$

and for the convective form of the mass transport equations

$$\begin{aligned}
\varepsilon s \acute{\mathfrak{R}}_k \frac{\partial C_k}{\partial t} + \boldsymbol{q} \cdot \nabla C_k - \nabla \cdot (\boldsymbol{D}_k \cdot \nabla C_k) + (\varepsilon s \vartheta_k \mathfrak{R}_k + Q_h) C_k &= \hat{R}_k + Q_{kw} + Q_k \\
&\quad \text{species } k \text{ of liquid phase } l \\
\varepsilon_s \frac{\partial C_k^s}{\partial t} + \varepsilon_s \vartheta_k C_k^s &= \hat{R}_k + Q_k \\
&\quad \text{species } k \text{ of solid phase } s
\end{aligned} \tag{12.2}$$

associated with the constitutive relations[1]

[1] In 3D Cartesian coordinates the components of the mechanical dispersion tensor $\boldsymbol{D}_{\text{mech}}$ for the classic Scheidegger-Bear dispersion model, cf. (3.182), are

$$\begin{aligned}
D_{\text{mech},11} &= \tfrac{1}{q}\left(\beta_L q_1^2 + \beta_T q_2^2 + \beta_T q_3^2\right) \\
D_{\text{mech},22} &= \tfrac{1}{q}\left(\beta_T q_1^2 + \beta_L q_2^2 + \beta_T q_3^2\right) \\
D_{\text{mech},33} &= \tfrac{1}{q}\left(\beta_T q_1^2 + \beta_T q_2^2 + \beta_L q_3^2\right) \\
D_{\text{mech},12} &= D_{\text{mech},21} = (\beta_L - \beta_T)\tfrac{q_1 q_2}{q} \\
D_{\text{mech},13} &= D_{\text{mech},31} = (\beta_L - \beta_T)\tfrac{q_1 q_3}{q} \\
D_{\text{mech},23} &= D_{\text{mech},32} = (\beta_L - \beta_T)\tfrac{q_2 q_3}{q}
\end{aligned}$$

where $\boldsymbol{q}^T = \left(q_1 \; q_2 \; q_3\right)$ and $q = \|\boldsymbol{q}\|$. In strictly stratified aquifer system, where the transverse dispersion in the vertical x_3-direction can be much smaller than in the horizontal, Burnett and Frind [65] proposed the 3D mechanical dispersion tensor in an alternative form

$$\begin{aligned}
D_{\text{mech},11} &= \tfrac{1}{q}\left(\beta_L q_1^2 + \beta_{TH} q_2^2 + \beta_{TV} q_3^2\right) \\
D_{\text{mech},22} &= \tfrac{1}{q}\left(\beta_{TH} q_1^2 + \beta_L q_2^2 + \beta_{TV} q_3^2\right) \\
D_{\text{mech},33} &= \tfrac{1}{q}\left(\beta_{TV} q_1^2 + \beta_{TV} q_2^2 + \beta_L q_3^2\right) \\
D_{\text{mech},12} &= D_{\text{mech},21} = (\beta_L - \beta_{TH})\tfrac{q_1 q_2}{q} \\
D_{\text{mech},13} &= D_{\text{mech},31} = (\beta_L - \beta_{TV})\tfrac{q_1 q_3}{q} \\
D_{\text{mech},23} &= D_{\text{mech},32} = (\beta_L - \beta_{TV})\tfrac{q_2 q_3}{q}
\end{aligned}$$

splitting the transverse dispersivity into a horizontal transverse dispersivity β_{TH} and a vertical transverse dispersivity β_{TV}, where it is assumed that $\beta_{TH} \gg \beta_{TV}$. However, as noted by Bear and Cheng [38], Burnett and Find's mechanical dispersion tensor is not consistent with the basic

$$
\begin{aligned}
\boldsymbol{D}_k &= \varepsilon s D_k \boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}} \\
\boldsymbol{D}_{\text{mech}} &= \beta_T \|\boldsymbol{q}\| \boldsymbol{\delta} + (\beta_L - \beta_T) \tfrac{\boldsymbol{q} \otimes \boldsymbol{q}}{\|\boldsymbol{q}\|} \\
\varepsilon_s &= 1 - \varepsilon \\
\mathfrak{R}_k &= 1 + \big(\tfrac{1-\varepsilon}{\varepsilon}\big) \varphi_k \\
\acute{\mathfrak{R}}_k &= 1 + \big(\tfrac{1-\varepsilon}{\varepsilon}\big) \tfrac{\partial(\varphi_k C_k)}{\partial C_k} \\
\varphi_k &= \begin{cases} \kappa_k & \text{Henry} \\ b_k^{\dagger} \, C_k^{b_k^{\ddagger}-1} & \text{Freundlich} \\ \frac{k_k^{\dagger}}{1+k_k^{\ddagger} C_k} & \text{Langmuir} \end{cases} \qquad \text{(Table 3.8)} \\
\hat{R}_k &= \hat{R}_k(\varepsilon, s, C_1^{\alpha}, \ldots, C_N^{\alpha}, T) \quad \alpha \in (l, s)
\end{aligned}
\tag{12.3}
$$

which has to be solved for species concentrations C_k, where for the sake of simplicity we drop the liquid phase index l and the species index k is considered unique in each phase (associated either with liquid l or solid s). In (12.1) and (12.2) a modified bulk reaction rate $\hat{R}_k$ is introduced, in which the linear decay reaction term is separated,[2] $\varepsilon_\alpha \vartheta_k C_k^{\alpha}$, $\alpha \in (l, s)$, and in addition the (non-reactive, zero-order) well-type SPC term Q_{kw} and a (non-reactive, zero-order) sink/source term are split off, where Q_{kw} is not applied to species of the solid phase. The reaction rate $\hat{R}_k$ is related to the previously defined bulk reaction rate R_k (5.95) and the deduced bulk reaction rate $\tilde{R}_k$ (5.96) via

$$
\begin{aligned}
R_k &= \tilde{R}_k - \textstyle\sum_\alpha \varepsilon_\alpha \vartheta_k C_k^{\alpha}, \ \alpha \in (l, s) \\
\tilde{R}_k &= \begin{cases} \hat{R}_k + Q_{kw} + Q_k & \text{species } k \text{ of liquid phase } l \\ \hat{R}_k + Q_k & \text{species } k \text{ of solid phase } s \end{cases}
\end{aligned}
\tag{12.4}
$$

Irreversible chemical reaction necessitates specification of rate expression R_k, actually $\hat{R}_k$, where polynomial representations in form of degradation and Arrhenius type kinetics or Monod type kinetics for more complex bio-chemical reaction systems are typical (see Sect. 5.5 for more). For variably saturated porous media additional constitutive relation exists for the saturation $s = s(\psi)$ as a function of pressure head ψ (cf. Sect. 10.2). Notice, in the given formulations of the mass transport equations we preferably use the linear Fick's law of hydrodynamic dispersion, (3.272) with $\Im_H = 0$. Non-Fickian dispersion is commonly related to variable-density problems as discussed in Chap. 11.

We note that usually there is no need to solve mass transport equations for all species N. Only species k of interest will be considered, which are important constituents of the chemical reaction process and/or have impacts on the flow

constitutive relations, such as derived in Sects. 3.8.5.4 and 3.8.5.5, and not conform with tensor transformation rules shown by Lichtner et al. [349].

[2] The separation of the linear decay term allows its numerically implicit treatment in the LHS of the resulting discrete equation system, while nonlinearities appearing in $\hat{R}_k$ require an appropriate iterative approach. Indeed, the reaction rate $\hat{R}_k$ can also incorporate a linear degradation term (in this case ϑ_k should be zero), however, its numerical computation can be less effective than in the direct separation.

and transport regime (e.g., spread and change of contaminants in an environmental flow system). Typically, the transport equations are specified for selected solutes[3](dissolved components) in the liquid phase l and sorbed species at the solid phase s, no more than the essential number of species $N^\star < N$ in total (cf. Sect. 3.9.2). In the above set of mass transport equations for species k occurring either in the liquid phase l or in the solid phase s, it is stipulated that any species k when also subjected to a sorptive equilibrium reaction (retardation) is referred to as a solute constituent in the liquid phase l, while a species exclusively associated with the solid phase s is deemed to be subjected to a (non-equilibrium) reaction kinetics.

The general species mass transport equations (12.1) or (12.2) have to be solved for C_k subject to a set of BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC (see Sect. 6.3.2), which is for the divergence form

$$\begin{aligned} C_k &= C_{kD} && \text{on} \quad \varGamma_{D_k} \times t[t_0, \infty) \\ (C_k \boldsymbol{q} - \boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= q^\dagger_{kC} && \text{on} \quad \varGamma_{N_k} \times t[t_0, \infty) \\ (C_k \boldsymbol{q} - \boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= -\varPhi^\dagger_{kC}(C_{kC} - C_k) && \text{on} \quad \varGamma_{C_k} \times t[t_0, \infty) \\ Q_{kw} &= -\textstyle\sum_w C_{kw} Q_w(t) \delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \varOmega \times t[t_0, \infty) \end{aligned} \qquad (12.5)$$

and for the convective form

$$\begin{aligned} C_k &= C_{kD} && \text{on} \quad \varGamma_{D_k} \times t[t_0, \infty) \\ -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= q_{kC} && \text{on} \quad \varGamma_{N_k} \times t[t_0, \infty) \\ -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= -\varPhi_{kC}(C_{kC} - C_k) && \text{on} \quad \varGamma_{C_k} \times t[t_0, \infty) \\ Q_{kw} &= -\textstyle\sum_w (C_{kw} - C_k) Q_w(t) \delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \varOmega \times t[t_0, \infty) \end{aligned} \qquad (12.6)$$

where the total boundary is $\varGamma = \varGamma_{C_k} \cup \varGamma_{N_k} \cup \varGamma_{C_k}$, $\forall k$. Note that there are no BC's for the species mass transport equations of the solid phase in (12.1) and

[3] In the special case of a *single-species solute*, where only one dissolved component exists, we can drop the species indicator k and write the governing mass transport equation (12.1) and (12.2), respectively, simply as

$$\begin{aligned} \tfrac{\partial}{\partial t}(\varepsilon s \mathfrak{R} C) + \nabla \cdot (\boldsymbol{q} C) - \nabla \cdot (\boldsymbol{D} \cdot \nabla C) + \varepsilon s \vartheta \mathfrak{R} C &= \hat{R} + Q_{Cw} + Q_C && \text{divergence form} \\ \varepsilon s \acute{\mathfrak{R}} \tfrac{\partial C}{\partial t} + \boldsymbol{q} \cdot \nabla C - \nabla \cdot (\boldsymbol{D} \cdot \nabla C) + (\varepsilon s \vartheta \mathfrak{R} + Q_h) C &= \hat{R} + Q_{Cw} + Q_C && \text{convective form} \end{aligned}$$

with

$$\begin{aligned} \boldsymbol{D} &= \varepsilon s D \boldsymbol{\delta} + \boldsymbol{D}_{\text{mech}} \\ \mathfrak{R} &= 1 + \left(\tfrac{1-\varepsilon}{\varepsilon}\right)\varphi \\ \acute{\mathfrak{R}} &= 1 + \left(\tfrac{1-\varepsilon}{\varepsilon}\right) \tfrac{\partial(\varphi C)}{\partial C} \\ \varphi &= \begin{cases} \kappa & \text{Henry} \\ b^\dagger\, C^{b^\ddagger - 1} & \text{Freundlich} \\ \frac{k^\dagger}{1 + k^\ddagger\, C} & \text{Langmuir} \end{cases} \qquad \text{(Table 3.8)} \\ \hat{R} &= \hat{R}(\varepsilon, s, C, T) \end{aligned}$$

for solving the solute concentration C associated with the liquid phase l, where Q_{Cw} and Q_C denote the well-type SPC term and the zero-order mass sink/source term, respectively, for the single-species solute.

(12.2). The normal mass fluxes on $\varGamma_{N_k}$ and $\varGamma_{C_k}$ differ between the divergence form and the convective form. As already discussed in Sects. 2.2.2 and 6.3.2 the divergence form imposes the total (advective plus dispersive) boundary mass flux, while the convective form imposes a dispersive mass flux at the boundary. However, the convective form can also be used to express a mass flux BC of an advective load by specifying the Cauchy-type BC in the form

$$-(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} = \underbrace{-\varPhi_{kC}}_{\boldsymbol{q}\cdot\boldsymbol{n}} (\underbrace{C_{kC}}_{\frac{q^{\dagger}_{kC}}{\boldsymbol{q}\cdot\boldsymbol{n}}} - C_k) \tag{12.7}$$

to obtain

$$(C_k \boldsymbol{q} - \boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} = q^{\dagger}_{kC} = (\boldsymbol{q} \cdot \boldsymbol{n}) C_{kC} \tag{12.8}$$

for a given advective normal boundary flux $\boldsymbol{q} \cdot \boldsymbol{n}$ and a boundary concentration C_{kC}, which is equivalent to a Neumann-type BC of the divergence form (cf. Sect. 6.3.2.3). Note further that OBC as discussed in Sect. 6.5.7 represents a special form of Neumann-type BC on $\varGamma_{N_{kO}} \subset \varGamma_{N_k} \subset \varGamma$, which will be treated either as a natural Neumann-type BC with $-(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} \approx 0$ or as implicit OBC (cf. Sect. 8.5.3).

The solution of the governing transient mass transport equations (12.1) and (12.2) requires IC in the form

$$C_k(\boldsymbol{x}, t_0) = C_{k,0}(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{12.9}$$

The essential parameters required for solving (12.1) and (12.2) with (12.5)–(12.9) are listed in Tables I.11 and I.13 of Appendix I. *Steady-state* mass transport conditions occur if $\partial C_k / \partial t$ approaches to zero.[4]

12.2.2 Horizontal 2D Problems

The governing equations for the essentially horizontal, vertically averaged species mass transport in unconfined and confined aquifers have been developed in Sect. 3.10.7 and summarized in Tables 3.10 and 3.11, respectively. The following 2D depth-integrated mass transport equations result

[4] Optionally, FEFLOW suppresses the time derivative term $\partial C_k / \partial t$ for solving steady-state solutions. A specific option exists, named *steady flow – transient transport*, in which the advective flow vector $\boldsymbol{q}$ is invariant with time.

$$
\begin{aligned}
&\frac{\partial}{\partial t}(\varepsilon\bar{\mathfrak{R}}_k C_k) + \nabla\cdot(\bar{\boldsymbol{q}}C_k) - \nabla\cdot(\bar{\boldsymbol{D}}_k\cdot\nabla C_k) + \varepsilon\vartheta_k\bar{\mathfrak{R}}_k C_k = \bar{\hat{R}}_k + \bar{Q}_{kw} + \bar{Q}_k \\
&\hspace{20em}\text{species } k \text{ of liquid phase } l \\
&\hspace{12em}\frac{\partial}{\partial t}(\varepsilon_s B C_k^s) + \varepsilon_s B\vartheta_k C_k^s = \bar{\hat{R}}_k + \bar{Q}_k \\
&\hspace{20em}\text{species } k \text{ of solid phase } s
\end{aligned}
\tag{12.10}
$$

written in the divergence form and

$$
\begin{aligned}
&\varepsilon\bar{\acute{\mathfrak{R}}}_k\frac{\partial C_k}{\partial t} + \bar{\boldsymbol{q}}\cdot\nabla C_k - \nabla\cdot(\bar{\boldsymbol{D}}_k\cdot\nabla C_k) + (\varepsilon\vartheta_k\bar{\mathfrak{R}}_k + \bar{Q}_h)C_k = \bar{\hat{R}}_k + \bar{Q}_{kw} + \bar{Q}_k \\
&\hspace{20em}\text{species } k \text{ of liquid phase } l \\
&\hspace{12em}\varepsilon_s B\frac{\partial C_k^s}{\partial t} + \varepsilon_s B\vartheta_k C_k^s = \bar{\hat{R}}_k + \bar{Q}_k \\
&\hspace{20em}\text{species } k \text{ of solid phase } s
\end{aligned}
\tag{12.11}
$$

written in the convective form, which are associated with the constitutive relations

$$
\begin{aligned}
B &= \begin{cases} h - f^B & \text{unconfined} \\ f^T - f^B & \text{confined} \end{cases} \\
\bar{\boldsymbol{D}}_k &= \varepsilon B D_k\boldsymbol{\delta} + \bar{\boldsymbol{D}}_{\text{mech}} \\
\bar{\boldsymbol{D}}_{\text{mech}} &= \beta_T\|\bar{\boldsymbol{q}}\|\boldsymbol{\delta} + (\beta_L - \beta_T)\tfrac{\bar{\boldsymbol{q}}\otimes\bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|} \\
\varepsilon_s &= 1 - \varepsilon \\
\bar{\mathfrak{R}}_k &= B\big[1 + \big(\tfrac{1-\varepsilon}{\varepsilon}\big)\varphi_k\big] \\
\bar{\acute{\mathfrak{R}}}_k &= B\big[1 + \big(\tfrac{1-\varepsilon}{\varepsilon}\big)\tfrac{\partial(\varphi_k C_k)}{\partial C_k}\big] \\
\varphi_k &= \begin{cases} \kappa_k & \text{Henry} \\ b_k^{\dagger}\,C_k^{b_k^{\ddagger}-1} & \text{Freundlich} \\ \frac{k_k^{\dagger}}{1+k_k^{\ddagger}C_k} & \text{Langmuir} \end{cases} \quad \text{(Table 3.8)} \\
\bar{\hat{R}}_k &= B\hat{R}_k \\
\hat{R}_k &= \hat{R}_k(\varepsilon, s, C_1^{\alpha}, \ldots, C_N^{\alpha}, T) \quad \alpha\in(l,s)
\end{aligned}
\tag{12.12}
$$

where similarly the deduced bulk reaction rate $\bar{\tilde{R}}_k = \bar{\hat{R}}_k + \bar{Q}_{kw} + \bar{Q}_k$ is suitably split into a depth-integrated bulk reaction rate $\bar{\hat{R}}_k$, a depth-integrated well-type SPC term $\bar{Q}_{kw}$ and a depth-integrated zero-order sink/source term for species k, where $\bar{Q}_{kw}$ is not applied to species in the solid phase. The solution of (12.10) or (12.11) for the species concentration C_k is associated with the following BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC

$$
\begin{aligned}
C_k &= C_{kD} && \text{on} \quad \varGamma_{D_k}\times t[t_0,\infty) \\
(C_k\bar{\boldsymbol{q}} - \bar{\boldsymbol{D}}_k\cdot\nabla C_k)\cdot\boldsymbol{n} &= \bar{q}_{kC}^{\dagger} && \text{on} \quad \varGamma_{N_k}\times t[t_0,\infty) \\
(C_k\bar{\boldsymbol{q}} - \bar{\boldsymbol{D}}_k\cdot\nabla C_k)\cdot\boldsymbol{n} &= -\bar{\Phi}_{kC}^{\dagger}(C_{kC} - C_k) && \text{on} \quad \varGamma_{C_k}\times t[t_0,\infty) \\
\bar{Q}_{kw} &= -\textstyle\sum_w C_{kw}Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w\in\varOmega\times t[t_0,\infty)
\end{aligned}
\tag{12.13}
$$

written for the divergence form of the mass transport equation and

$$\begin{aligned} C_k &= C_{kD} && \text{on} \quad \Gamma_{D_k} \times t[t_0, \infty) \\ -(\bar{\boldsymbol{D}}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= \bar{q}_{kC} && \text{on} \quad \Gamma_{N_k} \times t[t_0, \infty) \\ -(\bar{\boldsymbol{D}}_k \cdot \nabla C_k) \cdot \boldsymbol{n} &= -\bar{\Phi}_{kC}(C_{kC} - C_k) && \text{on} \quad \Gamma_{C_k} \times t[t_0, \infty) \\ \bar{Q}_{kw} &= -\textstyle\sum_w (C_{kw} - C_k) Q_w(t) \delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty) \end{aligned} \tag{12.14}$$

written for the convective form of the mass transport equation, imposed on $\Gamma = \Gamma_{C_k} \cup \Gamma_{N_k} \cup \Gamma_{C_k}$, $\forall k$, and with the IC of the form

$$C_k(\boldsymbol{x}, t_0) = C_{k,0}(\boldsymbol{x}) \quad \text{in} \quad \bar{\Omega} \tag{12.15}$$

The essential parameters required for solving (12.10) and (12.11) with (12.13), (12.14) and (12.15) are listed in Tables I.12 and I.14 of Appendix I.

12.3 Finite Element Formulation

In Chap. 8 the fundamental concepts of FEM are exemplified for an ADE of a scalar quantity, which is paradigmatic for the present species mass transport equations. Based on these principles given there we use now the GFEM to solve the governing mass transport equations (12.1) and (12.2) associated with the corresponding BC's (12.5), (12.6) and IC's (12.9). Since most of the details are equivalent to the ADE developments given in Chap. 8 we shall focus here only on aspects featuring the reactive multispecies mass transport. For convenience we restrict our developments to 3D, vertical 2D and axisymmetric mass transport problems (Sect. 12.2.1). The formulations for the horizontal 2D mass transport in unconfined and confined aquifers (Sect. 12.2.2) will appear rather similar and can be easily deduced from the given statements.

12.3.1 Weak Forms

According to Sect. 8.5 we can find analogously to the statements (8.48) and (8.55) the corresponding weak forms for the governing multispecies mass transport equation written in the divergence form (12.1) as

$$\int_\Omega w \frac{\partial(\varepsilon s \Re_k C_k)}{\partial t} d\Omega - \int_\Omega C_k \boldsymbol{q} \cdot \nabla w d\Omega + \int_\Omega \nabla w \cdot (\boldsymbol{D}_k \cdot \nabla C_k) d\Omega +$$
$$\int_\Omega w(\varepsilon s \vartheta_k \Re_k C_k - \hat{R}_k - Q_k) d\Omega + \sum_w w(\boldsymbol{x}_w) Q_w(t) C_{kw} +$$

$$\int_{\Gamma_{N_k}} wq_{kC}^{\dagger} d\Gamma - \int_{\Gamma_{C_k}} w\Phi_{kC}^{\dagger}(C_{kC} - C_k) d\Gamma = 0, \quad \forall w \in H_0^1(\Omega) \tag{12.16}$$

$$\int_{\Omega} w \frac{\partial(\varepsilon_s C_k^s)}{\partial t} d\Omega + \int_{\Omega} w(\varepsilon_s \vartheta_k C_k^s - \hat{R}_k - Q_k) d\Omega = 0, \quad \forall w \in H_0^1(\Omega) \tag{12.17}$$

and written in the convective form (12.2) as

$$\begin{aligned}
&\int_{\Omega} w\varepsilon s \acute{\Re}_k \frac{\partial C_k}{\partial t} d\Omega + \int_{\Omega} w\boldsymbol{q} \cdot \nabla C_k d\Omega + \int_{\Omega} \nabla w \cdot (\boldsymbol{D}_k \cdot \nabla C_k) d\Omega + \\
&\int_{\Omega} w[(\varepsilon s \vartheta_k \Re_k + Q_h) C_k - \hat{R}_k - Q_k] d\Omega + \sum_{w} w(\boldsymbol{x}_w) Q_w(t)(C_{kw} - C_k) + \\
&\int_{\Gamma_{N_k}} wq_{kC} d\Gamma - \int_{\Gamma_{C_k}} w\Phi_{kC}(C_{kC} - C_k) d\Gamma = 0, \quad \forall w \in H_0^1(\Omega)
\end{aligned} \tag{12.18}$$

$$\int_{\Omega} w\varepsilon_s \frac{\partial C_k}{\partial t} d\Omega + \int_{\Omega} w(\varepsilon_s \vartheta_k C_k^s - \hat{R}_k - Q_k) d\Omega = 0, \quad \forall w \in H_0^1(\Omega) \tag{12.19}$$

where w is a suitable weighting function and the boundary integrals are suitably separated into their segments $\Gamma = \Gamma_{D_k} \cup \Gamma_{N_k} \cup \Gamma_{C_k}$ imposed by the Dirichlet, Neumann and Cauchy-type BC's (12.5) and (12.6). OBC on $\Gamma_{N_{kO}} \subset \Gamma_{N_k}$ represents special implementations of Neumann-type BC.[5]

[5] A boundary with OBC on $\Gamma_{N_{kO}}$ can be separated from the Neumann boundary Γ_{N_k} so that for the divergence form

$$\int_{\Gamma_{N_k}} wq_{kC}^{\dagger} d\Gamma = \int_{\Gamma_{N_k} \setminus \Gamma_{N_{kO}}} wq_{kC}^{\dagger} d\Gamma + \int_{\Gamma_{N_{kO}}} w(C_k \boldsymbol{q} - \boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} d\Gamma$$

and for the convective form

$$\int_{\Gamma_{N_k}} wq_{kC} d\Gamma = \int_{\Gamma_{N_k} \setminus \Gamma_{N_{kO}}} wq_{kC} d\Gamma - \int_{\Gamma_{N_{kO}}} w(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} d\Gamma$$

The implicit treatment of OBC requires the incorporation of the $\Gamma_{N_{kO}}$—integrals into the LHS of the resulting matrix system (see below). In contrast, a natural Neumann-type BC with $-(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n} \approx 0$ on $\Gamma_{N_{kO}}$ is often the preferred alternative formulation for an OBC. Note, however, that for both cases in the divergence form the boundary flux $\boldsymbol{q} \cdot \boldsymbol{n}$ must be known a priori. The boundary flux $\boldsymbol{q} \cdot \boldsymbol{n}$ can be either explicitly given from a Neumann-type BC $q_h = \boldsymbol{q} \cdot \boldsymbol{n}$ for flow or must be computed by a postprocessing budget evaluation of the flow equation on the corresponding outflowing boundary section imposed by Dirichlet-type or Cauchy-type BC of flow.

12.3.2 GFEM and Resulting Nonlinear Matrix System

The weak statements (12.3.1)–(12.19) involve the unknown variable C_k of each species k occurring either on the liquid phase l or solid phase s. In using the FEM this variable is replaced by a *continuous approximation* that assumes the separability of space and time (see Sect. 8.4). Thus

$$C_k(\boldsymbol{x},t) \approx \sum_j N_j(\boldsymbol{x}) C_{kj}(t), \quad j = 1,\ldots,N_{\mathrm{P}},\ k = 1,\ldots,N \tag{12.20}$$

where j designates global nodal indices. Using the Galerkin method with the weighting function

$$w \to w_i = N_i, \quad i = 1,\ldots,N_{\mathrm{P}} \tag{12.21}$$

and applying the approximate solutions (12.20) in (12.3.1)–(12.19), we obtain the following matrix systems of each N_{P} equations (cf. Sect. 8.9) for each species k as follows

$$\boldsymbol{H}_k(\boldsymbol{C}) \cdot \dot{\boldsymbol{C}}_k + \boldsymbol{E}_k(\boldsymbol{C}) \cdot \boldsymbol{C}_k - \boldsymbol{R}_k(\boldsymbol{C}) = \boldsymbol{0} \quad (k = 1,\ldots,N) \tag{12.22}$$

or

$$\begin{aligned}
\boldsymbol{H}_1(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\dot{\boldsymbol{C}}_1 + \boldsymbol{E}_1(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\boldsymbol{C}_1 - \boldsymbol{R}_1(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N) &= \boldsymbol{0}\\
\boldsymbol{H}_2(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\dot{\boldsymbol{C}}_2 + \boldsymbol{E}_2(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\boldsymbol{C}_2 - \boldsymbol{R}_2(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N) &= \boldsymbol{0}\\
\boldsymbol{H}_3(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\dot{\boldsymbol{C}}_3 + \boldsymbol{E}_3(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\boldsymbol{C}_3 - \boldsymbol{R}_3(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N) &= \boldsymbol{0}\\
&\vdots\\
\boldsymbol{H}_N(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\dot{\boldsymbol{C}}_N + \boldsymbol{E}_N(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N)\cdot\boldsymbol{C}_N - \boldsymbol{R}_N(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_N) &= \boldsymbol{0}
\end{aligned} \tag{12.23}$$

and alternatively written in a compact form as

$$\boldsymbol{H}(\boldsymbol{C})\cdot\dot{\boldsymbol{C}} + \boldsymbol{E}(\boldsymbol{C})\cdot\boldsymbol{C} - \boldsymbol{R}(\boldsymbol{C}) = \boldsymbol{0} \tag{12.24}$$

or

$$\begin{pmatrix} \boldsymbol{H}_1 & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{H}_2 & \boldsymbol{0} & \ldots & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{H}_3 & \ldots & \boldsymbol{0}\\ & & & \ddots & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{H}_N \end{pmatrix} \cdot \begin{pmatrix} \dot{\boldsymbol{C}}_1\\ \dot{\boldsymbol{C}}_2\\ \dot{\boldsymbol{C}}_3\\ \vdots\\ \dot{\boldsymbol{C}}_N \end{pmatrix} + \begin{pmatrix} \boldsymbol{E}_1 & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{E}_2 & \boldsymbol{0} & \ldots & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{E}_3 & \ldots & \boldsymbol{0}\\ & & & \ddots & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{E}_N \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{C}_1\\ \boldsymbol{C}_2\\ \boldsymbol{C}_3\\ \vdots\\ \boldsymbol{C}_N \end{pmatrix} - \begin{pmatrix} \boldsymbol{R}_1\\ \boldsymbol{R}_2\\ \boldsymbol{R}_3\\ \vdots\\ \boldsymbol{R}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{0}\\ \boldsymbol{0}\\ \boldsymbol{0}\\ \vdots\\ \boldsymbol{0} \end{pmatrix} \tag{12.25}$$

with

$$\boldsymbol{C}_k = \begin{pmatrix} C_{k1} \\ C_{k2} \\ \vdots \\ C_{kN_\mathrm{P}} \end{pmatrix}, \quad \dot{\boldsymbol{C}}_k = \begin{pmatrix} \frac{dC_{k1}}{dt} \\ \frac{dC_{k2}}{dt} \\ \vdots \\ \frac{dC_{kN_\mathrm{P}}}{dt} \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_1 \\ \boldsymbol{C}_2 \\ \boldsymbol{C}_3 \\ \vdots \\ \boldsymbol{C}_N \end{pmatrix}, \quad \dot{\boldsymbol{C}} = \begin{pmatrix} \dot{\boldsymbol{C}}_1 \\ \dot{\boldsymbol{C}}_2 \\ \dot{\boldsymbol{C}}_3 \\ \vdots \\ \dot{\boldsymbol{C}}_N \end{pmatrix} \tag{12.26}$$

showing the major nonlinearities in parentheses, where the matrices and RHS vectors are given for species k of the liquid phase l as

$$\begin{aligned}
\boldsymbol{H}_k = H_{ij,k} &= \begin{cases} \sum_e \int_{\Omega^e} \varepsilon^e s^e \mathfrak{R}_k^e(\boldsymbol{C}^e)\, N_i N_j \, d\Omega^e & \text{divergence form} \\ \sum_e \int_{\Omega^e} \varepsilon^e s^e \acute{\mathfrak{R}}_k^e(\boldsymbol{C}^e)\, N_i N_j \, d\Omega^e & \text{convective form} \end{cases} \\
\boldsymbol{E}_k = E_{ij,k} &= \begin{cases} \sum_e \Big(-\int_{\Omega^e} \boldsymbol{q}^e \cdot \nabla N_i N_j \, d\Omega^e + \int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{D}_k^e \cdot \nabla N_j) d\Omega^e + \\ \quad \int_{\Omega^e} \big[\varepsilon^e s^e \vartheta_k^e \mathfrak{R}_k^e(\boldsymbol{C}^e) + \frac{\partial(\varepsilon^e s^e \mathfrak{R}_k^e(\boldsymbol{C}^e))}{\partial t}\big] N_i N_j \, d\Omega^e + \\ \quad \int_{\Gamma_{C_k}^e} \Phi_{kC}^{\dagger e} N_i N_j \, d\Gamma^e + \int_{\Gamma_{N_{kO}}^e} N_i (\boldsymbol{q}^e N_j - \boldsymbol{D}_k^e \cdot \nabla N_j) \cdot \boldsymbol{n} \, d\Gamma^e \Big) \\ & \text{divergence form} \\ \sum_e \Big(\int_{\Omega^e} N_i \boldsymbol{q}^e \cdot \nabla N_j \, d\Omega^e + \int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{D}_k^e \cdot \nabla N_j) d\Omega^e + \\ \quad \int_{\Omega^e} [\varepsilon^e s^e \vartheta_k^e \mathfrak{R}_k^e(\boldsymbol{C}^e) + Q_h^e] N_i N_j \, d\Omega^e + \\ \quad \int_{\Gamma_{C_k}^e} \Phi_{kC}^e N_i N_j \, d\Gamma^e - \int_{\Gamma_{N_{kO}}^e} N_i (\boldsymbol{D}_k^e \cdot \nabla N_j) \cdot \boldsymbol{n} \, d\Gamma^e \Big) - \delta_{ij} Q_w(t)\big|_i \\ & \text{convective form} \end{cases} \\
\boldsymbol{R}_k = R_{i,k} &= \begin{cases} \sum_e \Big(\int_{\Omega^e} N_i [\hat{R}_k^e(\boldsymbol{C}^e) + Q_k^e] d\Omega^e + \int_{\Gamma_{C_k}^e} N_i \Phi_{kC}^{\dagger e} C_{kC}^e \, d\Gamma^e - \\ \quad \int_{\Gamma_{N_k}^e \setminus \Gamma_{N_{kO}}^e} N_i q_{kC}^{\dagger e} \, d\Gamma^e \Big) - C_{kw} Q_w(t)\big|_i & \text{divergence form} \\ \sum_e \Big(\int_{\Omega^e} N_i [\hat{R}_k^e(\boldsymbol{C}^e) + Q_k^e] d\Omega^e + \int_{\Gamma_{C_k}^e} N_i \Phi_{kC}^e C_{kC}^e \, d\Gamma^e - \\ \quad \int_{\Gamma_{N_k}^e \setminus \Gamma_{N_{kO}}^e} N_i q_{kC}^e \, d\Gamma^e \Big) - C_{kw} Q_w(t)\big|_i & \text{convective form} \end{cases}
\end{aligned} \tag{12.27}$$

and for species k of the solid phase s as

$$\begin{aligned}
\boldsymbol{H}_k = H_{ij,k} &= \sum_e \int_{\Omega^e} \varepsilon_s^e N_i N_j \, d\Omega^e \\
\boldsymbol{E}_k = E_{ij,k} &= \begin{cases} \sum_e \int_{\Omega^e} \big(\varepsilon_s^e \vartheta_k^e + \frac{\partial \varepsilon_s^e}{\partial t}\big) N_i N_j \, d\Omega^e & \text{divergence form} \\ \sum_e \int_{\Omega^e} \varepsilon_s^e \vartheta_k^e N_i N_j \, d\Omega^e & \text{convective form} \end{cases} \\
\boldsymbol{R}_k = R_{i,k} &= \sum_e \int_{\Omega^e} N_i [\hat{R}_k^e(\boldsymbol{C}^e) + Q_k^e] d\Omega^e
\end{aligned} \tag{12.28}$$

in which $(i, j = 1, \ldots, N_{\mathrm{P}})$ and $(e = 1, \ldots, N_{\mathrm{E}})$, $\boldsymbol{H}_k = \boldsymbol{H}_k(\boldsymbol{C})$ $(k = 1, \ldots, N)$ are the nonlinear symmetric storage matrices including retardation effects for species k occurring in the liquid phase l, $\boldsymbol{E}_k = \boldsymbol{E}_k(\boldsymbol{C})$ $(k = 1, \ldots, N)$ are the unsymmetric 'conductance' matrices encompassing advection, dispersion and retardation effects for species k occurring in the liquid phase l as well as linear decay for species k in both the liquid phase l and the solid phase s, and $\boldsymbol{R}_k = \boldsymbol{R}_k(\boldsymbol{C})$ $(k = 1, \ldots, N)$ are the chemical rate vectors, which represent nonlinear dependencies on the total concentration vector $\boldsymbol{C}$ according to the considered reaction kinetics. We note that there is no advection and dispersion for species k belonging to the solid phase s. The integrals appearing in (12.27) and (12.28) are integrated on element level in the local coordinates as described in Sect. 8.12. Analytical evaluations of partial integral terms of (12.27) and (12.28) can be deduced from developments done in Appendix H for selected element types. The differential elements $d\Omega^e$ and $d\Gamma^e$ differ for 3D, 2D and axisymmetric problems as given by (8.122)–(8.124), respectively. Is is important to note that the resulting global system of equations (12.24) is *unsymmetric* since the matrix $\boldsymbol{E}$ is unsymmetric due to advection.

The matrix system (12.24) can be highly nonlinear mainly due to the dependence of the reaction rate vector $\boldsymbol{R}$ on $\boldsymbol{C}$ so that an efficient numerical solution strategy is required, in particular for reactive multispecies transport problems.[6] One possibility would be the solution of the coupled matrix system (12.24) in a direct and simultaneous manner. Although mathematically rigorous, practical implementation of that approach is limited and not generally applicable to large, geometrically complex and multidimensional problems because of the significant memory and/or computational burden. The size of the coefficient matrices $\boldsymbol{H}$ and $\boldsymbol{E}$ in the discretized system (12.24) grows as a product of the number of nodes N_{P} and the number of applied species N. In general, the direct approach involves solving a $N_{\mathrm{P}} \times N$ system of nonlinear equations at each time plane. Furthermore, the system for a simultaneous solution can be ill-conditioned due to the significantly different scales of the processes involved. Alternatively, in order to reduce the computational requirements, a *decoupled* (or split-operator) solution strategy is preferred, in which the species equations are solved sequentially by using efficient iteration techniques. Kanney et al. [299] discussed different strategies of such split-operator approaches. Among a variety of split-operator techniques the sequential iterative approach (SIA) have proven superior and powerful. In FEFLOW, we prefer an adaptive error-controlled SIA strategy which is based on an efficient predictor-corrector time-stepping technique. In contrast to a common SIA technique the transport equations with the reaction terms are solved in an adaptive full time interval using predictor solutions to linearize the nonlinear reaction terms. The overall iteration control is

[6] For single-species solute transport (12.24) and (12.22) reduce to simplified matrix system, where nonlinearities can only occur due to nonlinear retardation (Freundlich or Langmuir adsorption isotherms) and/or higher-order kinetic reactions, however, subjected to the same species of solute.

fully embedded in a time-marching strategy via a sophisticated error-based time-step adaptation.

For advective-dominant mass transport the discretized system (12.24) can be easily combined with upwind strategies as introduced in Sect. 8.14. Useful upwind strategies refer to the SU and FU methods (Sect. 8.14.3), SC method (Sect. 8.14.4) and PGLS method (Sect. 8.14.5), in which the tensor of mechanical dispersion $\boldsymbol{D}_{\text{mech}}$ as part of the hydrodynamic dispersion tensor $\boldsymbol{D}_k$ is appropriately modified by stabilization terms in dependence on the actual spatial and temporal discretizations or concentration gradients. The required modifications of $\boldsymbol{D}^e_{\text{mech}}$ for each element e were discussed in the preceding Sect. 11.6.3 and summarized in Table 11.3.

12.3.3 Adaptive SIA-Based Solution Strategy for Multispecies Mass Transport Embedded in the GLS Predictor-Corrector Time Integrator

The GLS predictor-corrector time integrator (Sect. 8.13.4) with automatically adapted time stepping has been shown very cost-efficient and robust for classes of nonlinear systems such as variably saturated problems (Sect. 10.7.5) and/or variable-density flow (Sect. 11.6.4). We also prefer this technique[7] for solving the present transient chemically reactive systems. At a multispecies presence ($N^{\star} > 1$)

[7] Alternatively to the GLS predictor-corrector method, the time integration of (12.22) for each species k by using the simple $\theta-$method (Sect. 8.13.4) gives

$$\begin{array}{l}\left(\frac{\boldsymbol{H}_k(\boldsymbol{C}_{n+1})}{\Delta t_n} + \boldsymbol{E}_k(\boldsymbol{C}_{n+1})\theta\right) \cdot \boldsymbol{C}_{k,n+1} = \\ \left(\frac{\boldsymbol{H}_k(\boldsymbol{C}_{n+1})}{\Delta t_n} - \boldsymbol{E}_k(\boldsymbol{C}_{n+1})(1-\theta)\right) \cdot \boldsymbol{C}_{k,n} + \big(\boldsymbol{R}_k(\boldsymbol{C}_{n+1})\theta + \boldsymbol{R}_k(\boldsymbol{C}_n)(1-\theta)\big)\end{array}$$

where $\theta \in (\frac{1}{2}, 1)$ for the Crank-Nicolson and the fully implicit scheme, respectively. For chemically reactive processes a nonlinear matrix system $\boldsymbol{R}^{\star}_{k,n+1} = \boldsymbol{A}_k(\boldsymbol{C}_{n+1}) \cdot \boldsymbol{C}_{k,n+1} - \boldsymbol{Z}_k(\boldsymbol{C}_{n+1}, \boldsymbol{C}_n) = \boldsymbol{0}$ results, which must be iteratively solved either via the Picard method (Sect. 8.18.1)

$$\boldsymbol{A}_k(\boldsymbol{C}^{\tau}_{n+1}) \cdot \boldsymbol{C}^{\tau+1}_{k,n+1} = \boldsymbol{Z}_k(\boldsymbol{C}^{\tau}_{n+1}, \boldsymbol{C}_n) \quad \tau = 0, 1, 2, \ldots$$

or via the Newton method (Sect. 8.18.2)

$$\begin{aligned} \boldsymbol{J}_k(\boldsymbol{C}^{\tau}_{n+1}) \cdot \Delta\boldsymbol{C}^{\tau}_{k,n+1} &= -\boldsymbol{R}^{\star}_{k,n+1}(\boldsymbol{C}^{\tau}_{n+1}, \boldsymbol{C}_n) \quad \tau = 0, 1, 2, \ldots \\ \Delta\boldsymbol{C}^{\tau}_{k,n+1} &= \boldsymbol{C}^{\tau+1}_{k,n+1} - \boldsymbol{C}^{\tau}_{k,n+1} \\ \boldsymbol{J}_k(\boldsymbol{C}^{\tau}_{n+1}) &= \frac{\partial \boldsymbol{R}^{\star}_{k,n+1}(\boldsymbol{C}^{\tau}_{n+1}, \boldsymbol{C}_n)}{\partial \boldsymbol{C}^{\tau}_{k,n+1}} \end{aligned}$$

until satisfactory convergence is achieved for the iterations τ at each given time stage $n + 1$. Note that this iterative solution strategy is also applicable to *steady-state* mass transport problems if setting $\theta = 1$ and $\Delta t_n \rightarrow \infty$.

the solution is performed in a decoupled manner, where each $k-$species nonlinear matrix system (12.22) is sequentially solved and appropriately linearized by using the adaptive predictor-corrector time-stepping strategy consisting of the following working steps:

STEP 0: Initialization

Computation of the initial acceleration vectors $\dot{\boldsymbol{C}}_{k,0}$ for time plane $n = 0$ (once per $k-$species equation)

$$\boldsymbol{H}_k(\boldsymbol{C}_0) \cdot \dot{\boldsymbol{C}}_{k,0} = -\boldsymbol{E}_k(\boldsymbol{C}_0) \cdot \boldsymbol{C}_{k,0} + \boldsymbol{R}_k(\boldsymbol{C}_0) \quad (k = 1, \dots, N) \tag{12.29}$$

and guessing an initial time step Δt_0. The initial systems (12.29) are solved with the initial concentration vector $\boldsymbol{C}_0^T = (\boldsymbol{C}_{1,0}\ \boldsymbol{C}_{2,0}\ \boldsymbol{C}_{3,0}\ \dots\ \boldsymbol{C}_{N,0})$ known by the IC's (12.9) for each species k. They need to be solved only once at initial time t_0.

STEP 1: Predictor solutions

Perform explicit predictor solutions for all species k by using the 1st-order accurate FE and 2nd-order accurate AB scheme, respectively,

$$\boldsymbol{C}^p_{k,n+1} = \begin{cases} \boldsymbol{C}_{k,n} + \Delta t_n \dot{\boldsymbol{C}}_{k,n} & \text{FE predictor} \\ \boldsymbol{C}_{k,n} + \frac{\Delta t_n}{2}\left[\left(2 + \frac{\Delta t_n}{\Delta t_{n-1}}\right)\dot{\boldsymbol{C}}_{k,n} - \frac{\Delta t_n}{\Delta t_{n-1}}\dot{\boldsymbol{C}}_{k,n-1}\right] & \text{AB predictor} \end{cases} \tag{12.30}$$

where the superposed p denotes the predictor values at the new time plane $n+1$. Note that, since $\dot{\boldsymbol{C}}_{k,n-1}$ is required, the AB formula cannot be applied before the second step ($n = 1$). The prediction has to be started with the FE scheme, where $\dot{\boldsymbol{C}}_{k,0}$ is available from (12.29).

STEP 2: Corrector solutions

Do corrector solutions for the nonlinear matrix system (12.22) of each species k via the TR or BE scheme by applying the predictor solution $\boldsymbol{C}^{p\,T}_{n+1} = (\boldsymbol{C}^{p\,T}_{1,n+1}, \boldsymbol{C}^{p\,T}_{2,n+1}, \boldsymbol{C}^{p\,T}_{3,n+1}, \dots, \boldsymbol{C}^{p\,T}_{N,n+1})$ from (12.30) to linearize the species equations as

$$\left(\frac{\boldsymbol{H}_k(\boldsymbol{C}^p_{n+1})}{\theta \Delta t_n} + \boldsymbol{E}_k(\boldsymbol{C}^p_{n+1})\right) \cdot \boldsymbol{C}_{k,n+1} = \\ \boldsymbol{H}_k(\boldsymbol{C}^p_{n+1}) \cdot \left[\frac{\boldsymbol{C}_{k,n}}{\theta \Delta t_n} + \left(\tfrac{1}{\theta} - 1\right)\dot{\boldsymbol{C}}_{k,n}\right] + \boldsymbol{R}_k(\boldsymbol{C}^p_{n+1}) \tag{12.31}$$

to determine the concentration $\boldsymbol{C}_{k,n+1}$ for each species k at the new time plane $n+1$, where $\theta \in (\frac{1}{2}, 1)$ for the TR and BE scheme, respectively.

STEP 3: Updated accelerations

Update the new acceleration vectors for each species k by inverting the FE and BE, respectively:

$$\dot{\boldsymbol{C}}_{k,n+1} = \begin{cases} \frac{\boldsymbol{C}_{k,n+1}-\boldsymbol{C}_{k,n}}{\Delta t_n} & \text{FE} \\ \left(2 - \frac{\Delta t_{n-1}}{\Delta t_n + \Delta t_{n-1}}\right)\left(\frac{\boldsymbol{C}_{k,n+1}-\boldsymbol{C}_{k,n}}{\Delta t_n}\right) - \left(\frac{\Delta t_n}{\Delta t_n + \Delta t_{n-1}}\right)\left(\frac{\boldsymbol{C}_{k,n}-\boldsymbol{C}_{k,n-1}}{\Delta t_{n-1}}\right) & \text{AB} \end{cases} \tag{12.32}$$

to obtain $\dot{\boldsymbol{C}}_{k,n+1}$ at the new time plane $n+1$.

STEP 4: Error estimation

Compute the LTE for the FE/BE and AB/TR scheme as a function of the corrector and predictor solutions for each species k in the form (cf. Table 8.7)

$$\boldsymbol{d}_{k,n+1} = \varphi(\boldsymbol{C}_{k,n+1} - \boldsymbol{C}^p_{k,n+1}) \tag{12.33}$$

with

$$\varphi = \begin{cases} \frac{1}{2} & \text{for FE/BE} \\ \frac{1}{3\left(1+\frac{\Delta t_{n-1}}{\Delta t_n}\right)} & \text{for AB/TR} \end{cases} \tag{12.34}$$

Suitable error norms are applied to the LTE vector $\boldsymbol{d}_{k,n+1}$ for each species k. Commonly, the weighted RMS L_2 error norm

$$\|\boldsymbol{d}_{k,n+1}\|_{L_2} = \left[\frac{1}{N_\text{P}}\left(\sum_{i=1}^{N_\text{P}} \left|\frac{d_{k,i,n+1}}{C_{k,\max,n+1}}\right|^2\right)\right]^{1/2} \tag{12.35}$$

and the maximum L_∞ error norm

$$\|\boldsymbol{d}_{k,n+1}\|_{L_\infty} = \frac{1}{C_{k,\max,n+1}} \max_i |d_{k,i,n+1}| \tag{12.36}$$

are chosen, where $C_{k,\max,n+1}$ corresponds to the maximum values of $k-$species concentration detected at the time plane $n+1$ and used to normalize the solution vector.

STEP 5: Tactic of time stepping and error control

Predict the potential new $k-$specific time-step lengths by means of the error estimate (12.33) for each species k, the current time step size Δt_n and a user-specified error tolerance ϵ as:

$$\Delta t_{k,n+1} = \Delta t_n \left(\frac{\epsilon}{\|\boldsymbol{d}_{k,n+1}\|_{L_p}}\right)^{1/\lambda} \tag{12.37}$$

where

$$\begin{aligned} \lambda &= \begin{cases} 2 & \text{for FE/BE} \\ 3 & \text{for AB/TR} \end{cases} \\ p &= \begin{cases} 2 & \text{for RMS error norm} \\ \infty & \text{for maximum error norm} \end{cases} \end{aligned} \tag{12.38}$$

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

(1) If

$$\Delta t_{k,n+1} \geq \Delta t_n \tag{12.39}$$

the solution $\boldsymbol{C}_{k,n+1}$ for the species equation k is accurate within the error bound defined by ϵ and the increase of the time step is always accepted.

(2) Else if

$$\gamma \Delta t_n \leq \Delta t_{k,n+1} < \Delta t_n \tag{12.40}$$

where γ is typically 0.85, the kth solution $\boldsymbol{C}_{k,n+1}$ is accepted but the time step is not changed, i.e., $\Delta t_{k,n+1} = \Delta t_n$.

(3) Else if

$$\Delta t_{k,n+1} < \gamma \Delta t_n \tag{12.41}$$

the solution $\boldsymbol{C}_{k,n+1}$ cannot be accepted within the required error tolerance ϵ and has to be rejected. The proposed new time step size (12.37) has to be reduced according to

$$\Delta t_{k,n+1}^{\text{red}} = \frac{\Delta t_n^2}{\Delta t_{n+1}} \left(\frac{\epsilon}{\|\boldsymbol{d}_{k,n+1}\|_{L_p}} \right)^{\varsigma} \quad (\varsigma = 1 \text{ for FE/BE and } \varsigma = 2/3 \text{ for AB/TR}) \tag{12.42}$$

and the solution of all species k must be repeated for the time plane $n+1$ with $\Delta t_n = \min_k (\Delta t_{k,n+1}^{\text{red}})$.

(4) If the criteria (12.39) and (12.40) are satisfied by all species equations and the solutions $\boldsymbol{C}_{k,n+1}$ can be accepted for all species k within the required error tolerance ϵ, the new time step is determined from the minimum of the $k-$specific time step lengths, viz.,

$$\Delta t_{n+1} = \min_k (\Delta t_{k,n+1}) \tag{12.43}$$

and the time stepping procedure proceeds to the new time plane $n+2$ with the time step Δt_{n+1} (12.43).

It is important to note that the error tolerance ϵ is the only user-specified parameter to control the entire nonlinear and transient 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., $\Delta t_{n+1} \leq \Delta t_{\max}$. Secondly, the rate for changing the time-step size $\Xi = \Delta t_{n+1}/\Delta t_n$ can also be limited, where $\Xi > 1$ can be 2, 3 or even more. Using these constraints the actually increased new time step results as $\Delta t_{n+1}^{\text{actual}} = \min(\Delta t_{n+1}, \Delta t_{\max}, \Xi \Delta t_n)$.

The predictor-corrector strategy fully monitors the nonlinear and transient solution process via the time LTE in which the size of the time step is cheaply and automatically varied in accordance with the overall accuracy requirements. The time step is increased whenever possible and decreased only when necessary. It is evident to note that by monitoring the temporal accuracy requirements, at the same time the solution strategy provides an efficient control of the nonlinearities of the species transport equation system via the predictor solutions. Due to the power of the predictor-corrector strategy any additional iterative feedback within the adapted time steps can be avoided.

12.4 Mass Budget Analysis

We use the CBFM, as introduced in Sect. 8.19.2, for obtaining a precise mass budget analysis. It is based on the specific weak formulations of the governing mass transport equations. The corresponding boundary mass fluxes on Γ have to be evaluated from the basic weak statements (12.3.1) and (12.18) of the divergence and convective form, respectively, written as

$$\begin{aligned}\int_\Gamma N_i\, q_{n_{kC}}^{\dagger}\, d\Gamma = -\int_\Omega N_i \frac{\partial(\varepsilon s \Re_k C_k)}{\partial t} d\Omega + \int_\Omega C_k \boldsymbol{q} \cdot \nabla N_i d\Omega - \\ \int_\Omega \nabla N_i \cdot (\boldsymbol{D}_k \cdot \nabla C_k) d\Omega - \int_\Omega N_i (\varepsilon s \vartheta_k \Re_k C_k - \hat{R}_k - Q_k) d\Omega - \\ C_{kw} Q_w(t)|_i \end{aligned} \tag{12.44}$$

$$\begin{aligned}\int_\Gamma N_i\, q_{n_{kC}}\, d\Gamma = -\int_\Omega N_i \varepsilon s \acute{\Re}_k \frac{\partial C_k}{\partial t} d\Omega - \int_\Omega N_i \boldsymbol{q} \cdot \nabla C_k d\Omega - \\ \int_\Omega \nabla N_i \cdot (\boldsymbol{D}_k \cdot \nabla C_k) d\Omega - \int_\Omega N_i [(\varepsilon s \vartheta_k \Re_k + Q_h) C_k - \hat{R}_k - Q_k] d\Omega - \\ (C_{kw} - C_k) Q_w(t)|_i \end{aligned} \tag{12.45}$$

to compute $q_{n_{kC}}^{\dagger}$ or $q_{n_{kC}}$, where C_k is known at evaluation time t_{n+1}. Note that the boundary mass flux $q_{n_{kC}}^{\dagger} = (C_k \boldsymbol{q} - \boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n}$ of the divergence form encompasses the total mass flux consisting of the advective and dispersive parts, while the boundary mass flux $q_{n_{kC}} = -(\boldsymbol{D}_k \cdot \nabla C_k) \cdot \boldsymbol{n}$ of the convective form consists only of the dispersive part. Thus, for the convective form we need an additional balance expression of the missing advective part $q^a_{n_{kC}} = C_k \boldsymbol{q} \cdot \boldsymbol{n}$ to obtain $q^\dagger_{n_{kC}} = q_{n_{kC}} + q^a_{n_{kC}}$. This is attained by using an auxiliary weak formulation applied to the governing flow equation (10.5) as described in Sect. 8.19.2.4. It yields

$$\begin{aligned}\int_\Gamma N_i\, q^a_{n_{kC}}\, d\Gamma = -\int_\Omega \nabla N_i \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] C_k d\Omega\, - \\ \int_\Omega N_i \nabla C_k \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega\, + \\ \int_\Omega N_i C_k (Q_h + Q_{hw} + Q_{\mathrm{EOB}}) d\Omega - \int_\Omega N_i C_k \Big(s S_o \frac{\partial h}{\partial t} + \varepsilon \frac{\partial s}{\partial t} \Big) d\Omega\end{aligned} \tag{12.46}$$

to compute $q^a_{n_{kC}}$, where h, s and C_k are known at evaluation time t_{n+1}. Expanding the boundary flux on Γ as described in Sect. 8.19.2 the following matrix system results to solve the consistent boundary total mass flux vector $\boldsymbol{q}^\dagger_{n_{kC}}$ for each species k, viz.,

$$\begin{aligned}\boldsymbol{M} \cdot \boldsymbol{q}^\dagger_{n_{kC}} &= -\boldsymbol{H}_k(\boldsymbol{C}) \cdot \dot{\boldsymbol{C}}_k - \boldsymbol{E}^\dagger_k(\boldsymbol{C}) \cdot \boldsymbol{C}_k + \boldsymbol{R}^\dagger_k(\boldsymbol{C}) \\ &\quad - \begin{cases} \boldsymbol{0} & \text{divergence form} \\ \boldsymbol{V}(\boldsymbol{h}) \cdot \boldsymbol{C}_k + \boldsymbol{A}(\boldsymbol{C}_k) \cdot \boldsymbol{h} - \boldsymbol{F}(\boldsymbol{C}_k, \boldsymbol{s}, \dot{\boldsymbol{h}}, \dot{\boldsymbol{s}}) & \text{convective form} \end{cases}\end{aligned} \tag{12.47}$$

for known $\boldsymbol{C}_k$, $\dot{\boldsymbol{C}}_k$, $\boldsymbol{h}$, $\dot{\boldsymbol{h}}$, $\boldsymbol{s}$ and $\dot{\boldsymbol{s}}$ at the corresponding evaluation time t_{n+1}, where $\boldsymbol{H}_k$ is defined in (12.27) and

$$\begin{aligned}\boldsymbol{M} = M_{ij} &= \textstyle\int_\Gamma N_i N_j d\Gamma \\ \boldsymbol{E}^\dagger_k = E^\dagger_{ij,k} &= \begin{cases} -\int_\Omega \boldsymbol{q} \cdot \nabla N_i N_j d\Omega + \int_\Omega \nabla N_i \cdot (\boldsymbol{D}_k \cdot \nabla N_j) d\Omega + \\ \int_\Omega (\varepsilon s \vartheta_k \Re_k + \frac{\partial(\varepsilon s \Re_k)}{\partial t}) N_i N_j d\Omega & \text{divergence form} \\ \int_\Omega N_i \boldsymbol{q} \cdot \nabla N_j d\Omega + \int_\Omega \nabla N_i \cdot (\boldsymbol{D}_k \cdot \nabla N_j) d\Omega + \\ \int_\Omega (\varepsilon s \vartheta_k \Re_k + Q_h) N_i N_j d\Omega - \delta_{ij} Q_w(t)\big|_i & \text{convective form} \end{cases} \\ \boldsymbol{R}^\dagger_k = R^\dagger_{i,k} &= \textstyle\int_\Omega N_i (\hat{R}_k + Q_k) d\Omega - C_{kw} Q_w(t)\big|_i \\ \boldsymbol{V} = V_{ij} &= \textstyle\int_\Omega N_i \nabla N_j \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega \\ \boldsymbol{A} = A_{ij} &= \textstyle\int_\Omega \nabla N_i \cdot [k_r \boldsymbol{K} f_\mu \cdot (\nabla N_j + \chi \boldsymbol{e})] C_k d\Omega \\ \boldsymbol{F} = F_i &= \textstyle\int_\Omega N_i C_k \big(Q_h + Q_{\mathrm{EOB}} - s S_o \frac{\partial h}{\partial t} - \varepsilon \frac{\partial s}{\partial t}\big) d\Omega - C_k Q_w(t)\big|_i\end{aligned} \tag{12.48}$$

in which $(i, j = 1, \ldots, N_{\mathrm{P}})$, $(e = 1, \ldots, N_{\mathrm{E}})$ and $(k = 1, \ldots, N)$. Note that $\boldsymbol{V}$, $\boldsymbol{A}$ and $\boldsymbol{F}$ are only needed for the convective form. In the budget analysis the integral boundary balance flux $\boldsymbol{Q}_{n_{kC}}$ is directly evaluated at each boundary node by

$$\begin{aligned}\boldsymbol{Q}_{n_{kC}} &= -\boldsymbol{M}\cdot\boldsymbol{q}^{\dagger}_{n_{kC}}\\ &= \boldsymbol{H}_k(\boldsymbol{C})\cdot\dot{\boldsymbol{C}}_k + \boldsymbol{E}^{\dagger}_k(\boldsymbol{C})\cdot\boldsymbol{C}_k - \boldsymbol{R}^{\dagger}_k(\boldsymbol{C})\\ &\quad+\begin{cases}\boldsymbol{0} & \text{divergence form}\\ \boldsymbol{V}(\boldsymbol{h})\cdot\boldsymbol{C}_k + \boldsymbol{A}(\boldsymbol{C}_k)\cdot\boldsymbol{h} - \boldsymbol{F}(\boldsymbol{C}_k,\boldsymbol{s},\dot{\boldsymbol{h}},\dot{\boldsymbol{s}}) & \text{convective form}\end{cases}\end{aligned} \tag{12.49}$$

where $\boldsymbol{Q}_{n_{kC}}$ corresponds to the nodal vector of the integral boundary mass flux.

12.5 Examples

12.5.1 Single-Species Solute Advective-Dispersive-Decay Transport in a Column

Considering a 1D column of homogeneous saturated porous medium in which a single-species solute intrudes with a constant concentration C_D, the flow in the column is maintained at a constant flux $q = \varepsilon v$ and in addition, the solute in the column continuously undergoes linear decay ϑ and linear adsorption $\Re$, then the governing mass transport equation (12.2) written in 1D $x-$coordinate reduces to

$$\Re\frac{\partial C}{\partial t} + v\frac{\partial C}{\partial x} - \mathcal{D}\frac{\partial^2 C}{\partial x^2} - \Re\vartheta C = 0 \tag{12.50}$$

with $v = \frac{q}{\varepsilon}$, $\mathcal{D} = D + \beta_L v$ and $\Re = 1 + (\frac{1-\varepsilon}{\varepsilon})\kappa$, for which the following analytical solution exists [33, 396, 540][8]

[8] Since the complementary error function erfc() is often in combination with exp(), it is numerically useful to introduce the function exf(a, b) defined as

$$\text{exf}(a,b) = \exp(a)\text{erfc}(b)$$

which is suitably approximated as follows [540]:

$$\text{exf}(a,b) \approx \begin{cases}\exp(a-b^2)(a_1\tau + a_2\tau^2 + a_3\tau^3 + a_4\tau^4 + a_5\tau^5) & \text{if } 0 \le b \le 3\\ \frac{1}{\sqrt{\pi}}\exp(a-b^2)/(b+0.5/(b+1/(b+1.5/(b+2/(b+2.5/(b+1)))))) & \text{if } b > 3\\ 2\exp(a) - \text{exf}(a,-b) & \text{if } b < 0\\ 0 \qquad \text{if } |a| > 170 \text{ and } b \le 0 \text{ or } |a-b^2| > 170 \text{ and } b > 0 & \end{cases}$$

where $\tau = 1/(1 + 0.3275911b)$ and $a_1 = 0.2548296$, $a_2 = -0.2844967$, $a_3 = 1.421414$, $a_4 = -1.453152$ and $a_5 = 1.061405$.

Table 12.1 Parameters and conditions used for the single-species solute advective-dispersive-decay transport in a column

Quantity	Symbol	Value	Unit
Column length	L	100	m
Constant flux	q	0.1	$\text{m}\,\text{d}^{-1}$
Porosity	ε	0.2	1
Constant velocity	$v = \frac{q}{\varepsilon}$	0.5	$\text{m}\,\text{d}^{-1}$
Henry sorption coefficient	κ	0.1	1
Retardation	$\Re = 1 + (\frac{1-\varepsilon}{\varepsilon})\kappa$	1.4	1
Decay rate	ϑ	$2 \cdot 10^{-8}$	s^{-1}
Molecular diffusion	D	0	$\text{m}^2\,\text{s}^{-1}$
Longitudinal dispersivity	β_L	0.1	m
Dispersion	$\mathcal{D} = D + \beta_L v$	$5.787 \cdot 10^{-7}$	$\text{m}^2\,\text{s}^{-1}$
IC and BC's			
Initial condition (IC) of C	C_0	0	$\text{mg}\,\text{l}^{-1}$
Dirichlet-type BC at $x = 0$	C_D	1	$\text{mg}\,\text{l}^{-1}$
Natural BC at $x = L$	$q_{n_C} = -\mathcal{D}\nabla C \cdot \boldsymbol{n}$	0	$\text{gm}^{-2}\,\text{d}^{-1}$
FEM			
Space increment	Δx	0.1	m
Initial time step size	Δt_0	10^{-4}	d
RMS error tolerance (AB/TR and FE/BE)	ϵ	10^{-4}	1
Simulation time period	t_{end}	200	d

$$C(x,t) = \tfrac{1}{2}C_D\Big[\exp\Big(\frac{(v-u)x}{2\mathcal{D}}\Big)\operatorname{erfc}\Big(\frac{\Re x - ut}{2\sqrt{\mathcal{D}\Re t}}\Big) + \exp\Big(\frac{(v+u)x}{2\mathcal{D}}\Big)\operatorname{erfc}\Big(\frac{\Re x + ut}{2\sqrt{\mathcal{D}\Re t}}\Big)\Big] \tag{12.51}$$

with

$$u = v\sqrt{1 + \frac{4\vartheta\Re\mathcal{D}}{v^2}}, \quad \operatorname{erfc}(a) = \tfrac{2}{\sqrt{\pi}}\int_a^\infty \exp(-\xi^2)d\xi \tag{12.52}$$

associated with the IC and BC's

$$C(x,0) = 0, \quad C(0,t) = C_D, \quad \tfrac{\partial C}{\partial x}(\infty,t) = 0 \tag{12.53}$$

to solve $C(x,t)$.

The analytical solution (12.51) will be compared with numerical results by using the parameters as summarized in Table 12.1. For the numerical simulations a uniform spatial discretization consisting of linear elements is used. The mesh is chosen sufficiently dense so that no upwinding is required. The adaptive GLS predictor-corrector time integrator is preferred, where both the 2nd-order accurate AB/TR and the 1st-order accurate FE/BE scheme are tested.

Fig. 12.1 Simulated versus analytical concentration profiles at different times t in days (single-species solute advective-dispersive-decay transport)

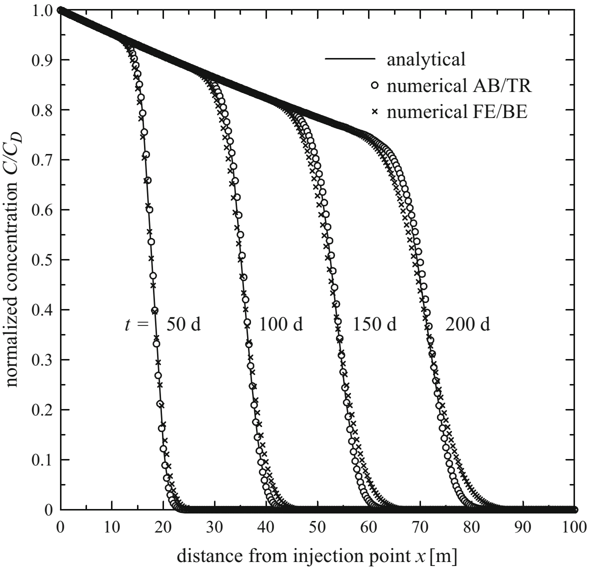
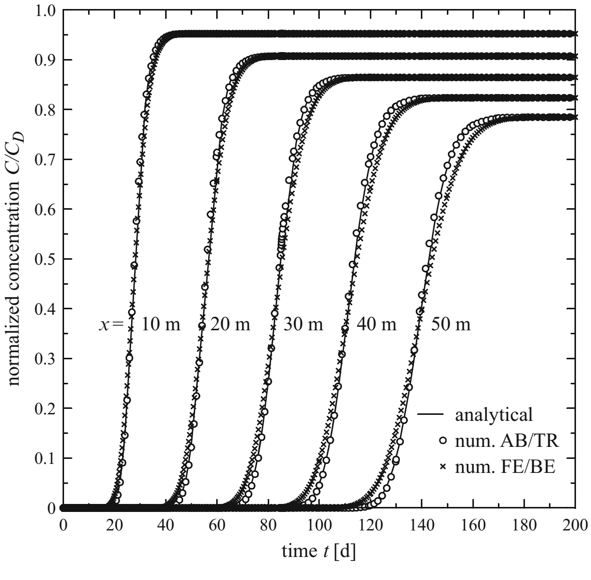


Fig. 12.2 Simulated versus analytical breakthrough curves at different distances x from injection point in meters (single-species solute advective-dispersive-decay transport)



The achieved numerical results in comparison with the analytical solutions are shown in Figs. 12.1 and 12.2. The agreements are rather well, in particular for the AB/TR scheme. Differences for the lower accurate FE/BE time stepping are revealed at later times, which indicate temporal discretization effects by numerical dispersion in the order $\mathcal{O}(\frac{\Delta t_n}{2}v^2)$ (cf. Sect. 8.15). The simulations over the period of

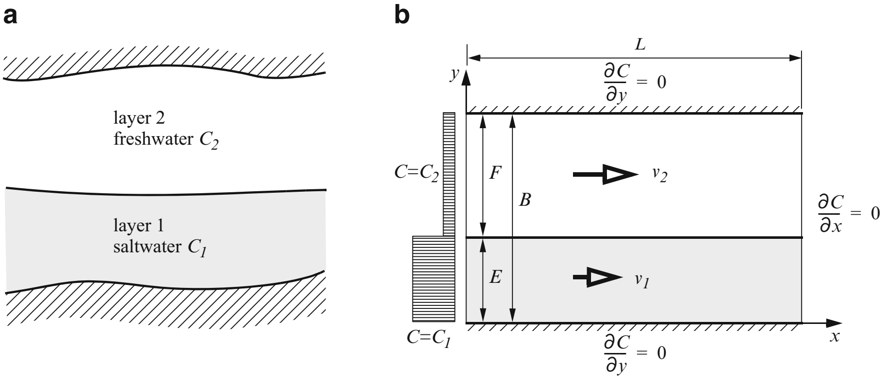

a
layer 2
freshwater C_2
layer 1
saltwater C_1
b
L
y
$\frac{\partial C}{\partial y} = 0$
$C=C_2$
F
B
v_2
$\frac{\partial C}{\partial x} = 0$
E
v_1
$C=C_1$
$\frac{\partial C}{\partial y} = 0$
x

Fig. 12.3 Paradigmatic bilayered aquifer structure: (**a**) principal undisturbed state and (**b**) schematic sketch (Modified from [513])

200 days required 210 time steps for the AB/TR scheme and 502 time steps for the FE/BE scheme.

12.5.2 Hydrodispersive Mixing of Single-Species Solute in a Bilayered Aquifer

Thiele and Diersch [513] studied a principal paradigmatic problem as illustrated in Fig. 12.3 for a confined alluvial aquifer having a bilayered structure, where the upper layer contains groundwater with lower salinity (freshwater) while in the underlying layer saline groundwater (saltwater) occurs. The major objective is to analyze the mechanism of transverse mixing the freshwater and saltwater flow under uniform (and possibly different) velocities in the two layers. Analytical solutions given by Thiele and Diersch [513] can be used to verify computational results when neglecting density effects and excluding chemical reaction.

For the present problem the governing mass transport equation (12.2) written in 2D $x - y-$coordinates reduces to

$$\Re \frac{\partial C}{\partial t} + v_x \frac{\partial C}{\partial x} - \mathcal{D}_{xx} \frac{\partial^2 C}{\partial x^2} - \mathcal{D}_{yy} \frac{\partial^2 C}{\partial y^2} = 0 \tag{12.54}$$

with $v_x = \frac{q_x}{\varepsilon}$, $\mathcal{D}_{xx} = D + \beta_L v_x$ and $\mathcal{D}_{yy} = D + \beta_T v_x$, where $C = C(x, y, t)$ represents the concentration of the single-species solute (salinity) to be solved. Note that $\mathcal{D}_{xy} = \mathcal{D}_{yx} = 0$ since $v_y = 0$ assuming an ideally $x-$parallel flow in the aquifer layers. Imposing the following IC and BC's *inhomogeneous IC:*

$$C(x,y,0) = \begin{cases} C_1 & (0 \le y \le E) \\ C_2 & (E < y \le B) \end{cases}, \quad (0 \le x \le L) \tag{12.55}$$

BC's:

$$C(0,y,t) = \begin{cases} C_1 & (0 \le y \le E) \\ C_2 & (E < y \le B) \end{cases} \tag{12.56}$$

$$\frac{\partial}{\partial x}C(L,y,t) = 0,\ \frac{\partial}{\partial y}C(x,0,t) = 0,\ \frac{\partial}{\partial y}C(x,B,t) = 0\ \ (0 \le x \le L,\ 0 \le y \le B) \tag{12.57}$$

an analytical solution for the case of uniform velocity $v = v_x = v_1 = v_2$ and neglected molecular diffusion $D = 0$ can be derived [513]:

$$\frac{C(x,y,t)-C_2}{C_1-C_2} = \frac{E}{B} + \sum_{i=1}^{\infty}\Big\{\frac{1}{i\pi}\sin\big(i\pi\frac{E}{B}\big)\cos\big(i\pi\frac{y}{B}\big)\Big[\exp\Big(\frac{x}{2\beta_L}(1-I_i)\Big)\mathrm{erfc}\Big(\frac{x-I_i vt}{2\sqrt{v\beta_L t}}\Big) + \exp\Big(\frac{x}{2\beta_L}(1+I_i)\Big)\mathrm{erfc}\Big(\frac{x+I_i vt}{2\sqrt{v\beta_L t}}\Big)\Big] + \frac{2}{i\pi}\sin\big(i\pi\frac{E}{B}\big)\cos\big(i\pi\frac{y}{B}\big)\exp\Big(-i^2\pi^2\frac{v\beta_T t}{B^2}\Big)\times \Big[1-\tfrac{1}{2}\mathrm{erfc}\Big(\frac{x-I_i vt}{2\sqrt{v\beta_L t}}\Big) - \tfrac{1}{2}\exp\Big(\frac{x}{\beta_L}\Big)\mathrm{erfc}\Big(\frac{x+I_i vt}{2\sqrt{v\beta_L t}}\Big)\Big]\Big\} \tag{12.58}$$

with

$$I_i = \sqrt{1 + \frac{4i^2\pi^2}{B^2}\beta_L\beta_T} \tag{12.59}$$

In using a homogeneous IC in form of $C(x,y,0) = C_2\ (0 \le x \le L,\ 0 \le y \le B)$ different to (12.55), (12.58) reduces to the Bruch and Street's analytical solution [60]

$$\frac{C(x,y,t)-C_2}{C_1-C_2} = \frac{E}{2B}\Big[\mathrm{erfc}\Big(\frac{x-vt}{2\sqrt{v\beta_L t}}\Big) + \exp\Big(\frac{x}{\beta_L}\Big)\mathrm{erfc}\Big(\frac{x+vt}{2\sqrt{v\beta_L t}}\Big)\Big] + \sum_{i=1}^{\infty}\Big\{\frac{1}{i\pi}\sin\big(i\pi\frac{E}{B}\big)\cos\big(i\pi\frac{y}{B}\big)\Big[\exp\Big(\frac{x}{2\beta_L}(1-I_i)\Big)\mathrm{erfc}\Big(\frac{x-I_i vt}{2\sqrt{v\beta_L t}}\Big) + \exp\Big(\frac{x}{2\beta_L}(1+I_i)\Big)\mathrm{erfc}\Big(\frac{x+I_i vt}{2\sqrt{v\beta_L t}}\Big)\Big]\Big\} \tag{12.60}$$

More complex analytical solution is given by Thiele and Diersch [513] for nonuniform $x-$parallel velocities $v_1 \neq v_2$ occurring in the two layers. Note that for evaluating the analytical exp(.)erfc(.) expressions appearing in (12.58) and (12.60) the more suitable exf(., .) function is used as already introduced in Sect. 12.5.1.

Table 12.2 Parameters and conditions used for the bilayered aquifer problem

Quantity	Symbol	Value	Unit
Domain shown in Fig. 12.3 with settings			
Domain length	L	3,500	m
Aquifer thickness	B	30	m
Layer 1 thickness	E	10	m
Layer 2 thickness	F	20	m
Constant horizontal flux	$q = q_x$	0.15	$\mathrm{m\,d^{-1}}$
Porosity	ε	0.3	1
Constant horizontal velocity	$v = \frac{q}{\varepsilon}$	0.5	$\mathrm{m\,d^{-1}}$
Retardation	$\Re$	1	1
Decay rate	ϑ	0	$\mathrm{s^{-1}}$
Molecular diffusion	D	0	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	5	m
Transverse dispersivity	β_T	0.5	m
IC and BC			
Initial condition (IC) of C (12.55)	$C_0 = \begin{cases} C_1 & (0 \le y \le E) \\ C_2 & (E < y \le B) \end{cases}$	$\begin{cases} 1 \\ 0 \end{cases}$	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC at $x = 0$ (12.56)	$C_D = \begin{cases} C_1 & (0 \le y \le E) \\ C_2 & (E < y \le B) \end{cases}$	$\begin{cases} 1 \\ 0 \end{cases}$	$\mathrm{mg\,l^{-1}}$
FEM			
Nonuniform 2D mesh of 1,000 × 70 linear and quadratic quadrilateral elements, GFEM			
Initial time step size	Δt_0	10^{-4}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	2,400	d

To compare the numerical results to the analytical solution (12.59) we simulate the mixing process in the bilayered aquifer for the case of a constant uniform (steady-state) velocity $v = v_x = v_1 = v_2$ with parameters and conditions as listed in Table 12.2. Unspecified BC's represent boundaries, at which natural BC's are imposed (12.57). The 2D cross-sectional model domain is appropriately discretized by $1{,}000 \times 70$ quadrilateral elements of both linear and quadratic element type. The structured mesh is uniformly discretized in $x-$direction $\Delta x = L/1{,}000$, however, in $y-$direction at the layer contact the element thickness Δy is about 4 cm and gradually increases with the distance from the interface. The computations are performed with GFEM without any upwind and the adaptive GLS AB/TR predictor-corrector time stepping. For the matrix equation solution the direct Gaussian elimination method is preferred.

The simulated breakthrough behavior in comparison to the analytical solution (12.58) is shown in Fig. 12.4 for two different points located in the upper layer 2. It reveals a typical nonmonothonic 'overshooting' characteristic resulting from the inhomogeneous IC (12.55) in which the lower layer 1 is fully pre-salinated at initial time. Those overshooting effects are thoroughly studied by Thiele and Diersch [513], even for nonuniform velocities in the aquifer layers. As shown in Fig. 12.4 the agreement with the analytical solution is rather well both for linear elements and

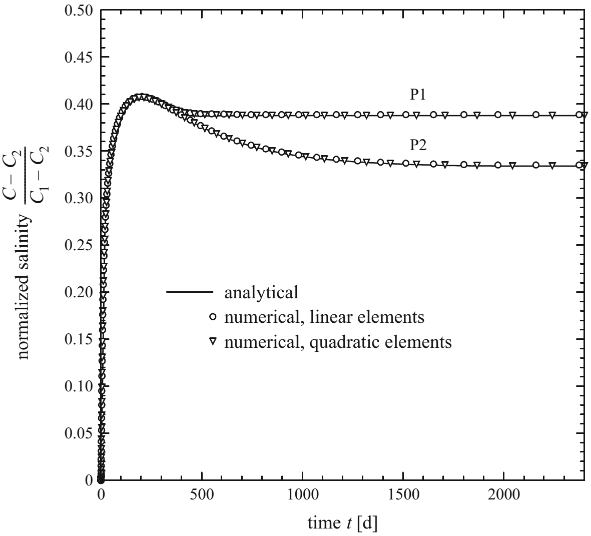


Fig. 12.4 Simulated versus analytical breakthrough curves of salinity $\frac{C-C_2}{C_1-C_2}$ at two different points P1$(x, y) = (200\,\text{m}, 12\,\text{m})$ and P2$(x, y) = (1{,}000\,\text{m}, 12\,\text{m})$ in a bilayered confined aquifer with uniform velocity $v = v_1 = v_2$ and inhomogeneous IC (12.55)

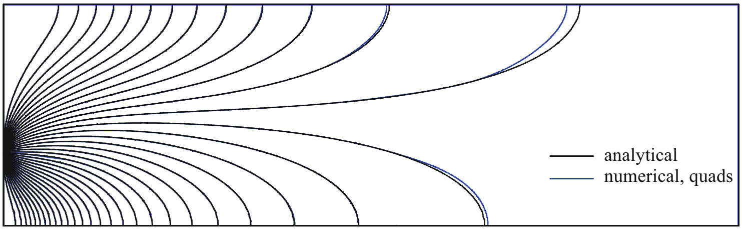


Fig. 12.5 Simulated salinity contours $\frac{C-C_2}{C_1-C_2}$ at $t = 2{,}400$ d in comparison to the analytical distribution (vertical exaggeration 10:1, shown domain ranges $0 \leq x \leq 1{,}000\,\text{m}$, $0 \leq y \leq B = 30\,\text{m}$) using inhomogeneous IC (12.55). Simulation results based on quadratic element mesh. Used contouring interval of normalized salinity is 0.025

somewhat better for quadratic elements. This is also illustrated in Fig. 12.5 for the simulated salinity contours in comparison to the analytical distribution. It indicates the fully mixing of salinity in the two layers for large distances x and elapsed times t approaching to a value of $(E\,C_1 + F\,C_2)/B = 0.3333$.

To expose the breakthrough behavior in contrast to the pre-salinated state of layer 1, Fig. 12.6 exhibits the salinity history at two points in layer 2 for the case in which both layers are filled by freshwater from beginning. Now, it reveals a monothonic increase of salinity without any overshoots. The attained numerical findings are shown in good agreement with the analytical solution, which is in this case given by Bruch and Street's expression (12.60).

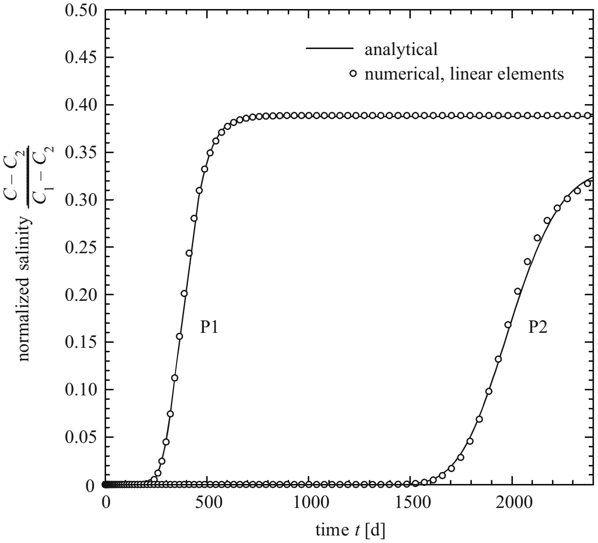


Fig. 12.6 Simulated versus analytical breakthrough curves of salinity $\frac{C-C_2}{C_1-C_2}$ at two different points P1(x, y) = (200 m, 12 m) and P2(x, y) = (1,000 m, 12 m) in a bilayered confined aquifer with uniform velocity $v = v_1 = v_2$ and homogeneous IC: $C_0 = C_2 = 0$

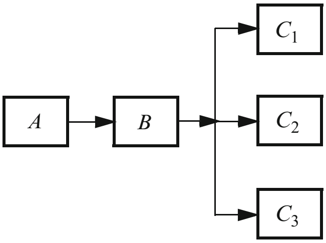


Fig. 12.7 Serial-parallel reaction network by Sun et al. [504]

12.5.3 *Multispecies Mass Transport with Comparison to Analytical Solutions*

12.5.3.1 Sun et al.'s 1D Serial-Parallel Reaction Problem

Sun et al. [504] present analytical solutions for 1D multispecies transport problems with serial and parallel reaction kinetics. As an example, the following reaction network is considered (Fig. 12.7) consisting of five essential species ($N^\star = 5$). The species B has three daughter species C_1, C_2, C_3. The reaction network of Fig. 12.7 can be decomposed into three serial reaction chains: $A \to B \to C_1$, $A \to B \to C_2$ and $A \to B \to C_3$. Accordingly, the following system of transport equations is considered:

Table 12.3 Problem parameters used for the 1D multispecies mass transport with serial-parallel reactions

Quantity	Symbol	Value	Unit
Extended length of column	$2L$	80	m
Longitudinal dispersivity	β_L	10	m
Pore velocity	v	0.4	$\mathrm{m\,d^{-1}}$
Molecular diffusion	D	0	$\mathrm{m^2\,s^{-1}}$
Dispersion	$\mathcal{D} = D + \beta_L v$	$4.63 \cdot 10^{-5}$	$\mathrm{m^2\,s^{-1}}$
Rate constant of species A	k_A	0.2	$\mathrm{d^{-1}}$
Rate constant of species B	k_B	0.1	$\mathrm{d^{-1}}$
Rate constant of species C_1	k_{C_1}	0.02	$\mathrm{d^{-1}}$
Rate constant of species C_2	k_{C_2}	0.02	$\mathrm{d^{-1}}$
Rate constant of species C_3	k_{C_3}	0.02	$\mathrm{d^{-1}}$
Stoichiometric coefficient of $A \to B$	ν_B	0.5	1
Stoichiometric coefficient of $B \to C_1$	ν_{C_1}	0.3	1
Stoichiometric coefficient of $B \to C_2$	ν_{C_2}	0.2	1
Stoichiometric coefficient of $B \to C_3$	ν_{C_3}	0.1	1
IC's and BC's			
Initial condition (IC) of C_k $(k = A, B, C_1, C_2, C_3)$	$C_{k,0}$	0	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_A at $x = 0$	C_{AD}	1	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_k at $x = 0$ $(k = B, C_1, C_2, C_3)$	C_{kD}	0	$\mathrm{mg\,l^{-1}}$
Natural BC for species C_k at $x = 2L$ $(k = A, B, C_1, C_2, C_3)$	$q_{n_{kC}} = -\mathcal{D}\nabla C_k \cdot \boldsymbol{n}$	0	$\mathrm{gm^{-2}\,d^{-1}}$
FEM			
Space increment	$\Delta x = 2L/600$	0.13333	m
Initial time step size	Δt_0	10^{-3}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	40	d

$$\begin{aligned}
\frac{\partial C_A}{\partial t} - \mathcal{D}\frac{\partial^2 C_A}{\partial x^2} + v\frac{\partial C_A}{\partial x} &= -k_A\, C_A \\
\frac{\partial C_B}{\partial t} - \mathcal{D}\frac{\partial^2 C_B}{\partial x^2} + v\frac{\partial C_B}{\partial x} &= \nu_B\, k_A\, C_A - k_B\, C_B \\
\frac{\partial C_{C_1}}{\partial t} - \mathcal{D}\frac{\partial^2 C_{C_1}}{\partial x^2} + v\frac{\partial C_{C_1}}{\partial x} &= \nu_{C_1}\, k_B\, C_B - k_{C_1}\, C_{C_1} \\
\frac{\partial C_{C_2}}{\partial t} - \mathcal{D}\frac{\partial^2 C_{C_2}}{\partial x^2} + v\frac{\partial C_{C_2}}{\partial x} &= \nu_{C_2}\, k_B\, C_B - k_{C_2}\, C_{C_2} \\
\frac{\partial C_{C_3}}{\partial t} - \mathcal{D}\frac{\partial^2 C_{C_3}}{\partial x^2} + v\frac{\partial C_{C_3}}{\partial x} &= \nu_{C_3}\, k_B\, C_B - k_{C_3}\, C_{C_3}
\end{aligned} \tag{12.61}$$

The RHS's of (12.61) represent the reaction rates R_A, R_B, R_{C_1}, R_{C_2} and R_{C_3}. In (12.61) $\mathcal{D} = D + \beta_L\, v$ is a constant dispersion coefficient, x is the 1D coordinate, v is a constant pore velocity, k_k $(k = A, B, C_1, C_2, C_2)$ are 1st-order rate constants and ν_k are corresponding stoichiometric coefficients. All species are diluted chemicals in a mobile liquid phase. The transport parameters assumed for this problem are listed in Table 12.3. There is no need to specify the porosity ε.

To simulate the 5-species transport problem the column of its double extent $2L$ is uniformly discretized by 600 linear elements. The automatic AB/TR time stepping procedure is applied. The simulation results can be compared to the analytical solution presented by Sun et al. [504]. Suppose the following IC's and BC's

$$\begin{array}{rll} C_k(x,0) = 0 & (k = A, B, C_1, C_2, C_2) & x \geq 0 \\ C_A(0,t) = 1 & & t > 0 \\ C_k(0,t) = 0 & (k = B, C_1, C_2, C_2) & t > 0 \\ C_k(\infty,t) = 0 & (k = A, B, C_1, C_2, C_2) & t > 0 \end{array} \tag{12.62}$$

the basic equation system (12.61) rewritten for the species k in the form

$$\mathcal{L}(C_k) = \nu_k\, k_{k-1}\, C_{k-1} - k_k\, C_k \tag{12.63}$$

where $\mathcal{L}(.)$ represents the differential operator, can be transformed by introducing the auxiliary variable a_k defined as

$$a_k = C_k + \begin{cases} 0 & k = 1 \\ \displaystyle\sum_{j=1}^{k-1} \prod_{i=j}^{k-1} \frac{\nu_{i+1}\, k_i}{k_i - k_k} C_j & k > 1 \end{cases} \tag{12.64}$$

to obtain the reactive transport equations in terms of a_k, viz.,

$$\mathcal{L}(a_k) = -k_k\, a_k \quad \forall k = 1, 2, \ldots, N^\star \tag{12.65}$$

Note that for $k = 1$ the transport equation (12.65) in terms of the first auxiliary variable is identical to the original equation (12.63) since $a_k = C_k$. The substituted equations (12.65) can be easily solved by the basic analytical formula

$$a_k(x,t) = \frac{a_{k0}}{2} \exp\Big(\frac{vx}{2\mathcal{D}}\Big)\Bigg[\exp(-u_k x)\mathrm{erfc}\Bigg(\frac{x - t\sqrt{v^2 + 4k_k\mathcal{D}}}{2\sqrt{\mathcal{D}t}}\Bigg) + \exp(u_k x)\mathrm{erfc}\Bigg(\frac{x + t\sqrt{v^2 + 4k_k\mathcal{D}}}{2\sqrt{\mathcal{D}t}}\Bigg)\Bigg] \tag{12.66}$$

where

$$u_k = \sqrt{\frac{v^2}{4\mathcal{D}^2} + \frac{k_k}{\mathcal{D}}} \tag{12.67}$$

and a_{k0} is the IC in terms of the auxiliary variable. The solutions of all concentrations C_k in the real untransformed domain can be determined by a successive substitution process using (12.64) in a reverse way

Table 12.4 Species ID's used in Sun et al.'s 1D problem

ID (= k)	Phase	Name
1	Liquid	A
2	Liquid	B
3	Liquid	C_1
4	Liquid	C_2
5	Liquid	C_3

$$C_k = a_k - \begin{cases} 0 & k = 1 \\ \displaystyle\sum_{j=1}^{k-1} \prod_{i=j}^{k-1} \frac{\nu_{i+1}\, k_i}{k_i - k_k} C_j & k > 1 \end{cases} \qquad (12.68)$$

where a_k is the solution from (12.66). Note that for evaluating the analytical exp(.)erfc(.) expressions in (12.66) the more suitable exf(., .) function is applied as already introduced in Sect. 12.5.1.

The infinite BC in (12.62) used in the analytical solution cannot be applied in the numerical context and is replaced by a natural Neumann-type BC imposed at the outlet boundary section of the double length $2L$ of the column, cf. Table 12.3. The reaction kinetics for the present problem is of a *degradation type*. We employ FEFLOW's reaction kinetics editor (see Sect. 5.5.4) to specify the reaction rates R_k as follows:

$$\begin{aligned} R_1 &= -\text{Rate}_1 \cdot C_1 \\ R_2 &= 0.5 \cdot \text{Rate}_1 \cdot C_1 - \text{Rate}_2 \cdot C_2 \\ R_3 &= 0.3 \cdot \text{Rate}_2 \cdot C_2 - \text{Rate}_3 \cdot C_3 \\ R_4 &= 0.2 \cdot \text{Rate}_2 \cdot C_2 - \text{Rate}_4 \cdot C_4 \\ R_5 &= 0.1 \cdot \text{Rate}_2 \cdot C_2 - \text{Rate}_5 \cdot C_5 \end{aligned} \qquad (12.69)$$

In (12.69) the parameters Rate$_k$ represent the reaction constants k_k of Table 12.3. The used species ID's are linked to the species names and phases as summarized in Table 12.4. A comparison of the computational results with the analytical solutions gives perfect agreements as exhibited in Fig. 12.8. The simulation by using the parameters and conditions as listed in Table 12.3 takes 72 time steps of variable length.

12.5.3.2 Sun et al.'s 3D First-Order Degradation Reaction Kinetics Problem

Sun et al. [503] have extended their analytical approach to 3D problems for homogeneous parameters and steady-state flow regimes. The solutions are demonstrated for a four-species transport in a 3D aquifer of $L \times D \times B = 100\,\text{m} \times 41\,\text{m} \times 25\,\text{m}$. For the finite-element analysis the symmetric half-domain is discretized by $120 \times 37 \times 50 = 222{,}000$ linear brick (hexahedral) elements consisting of 234,498 nodes as shown in Fig. 12.9.

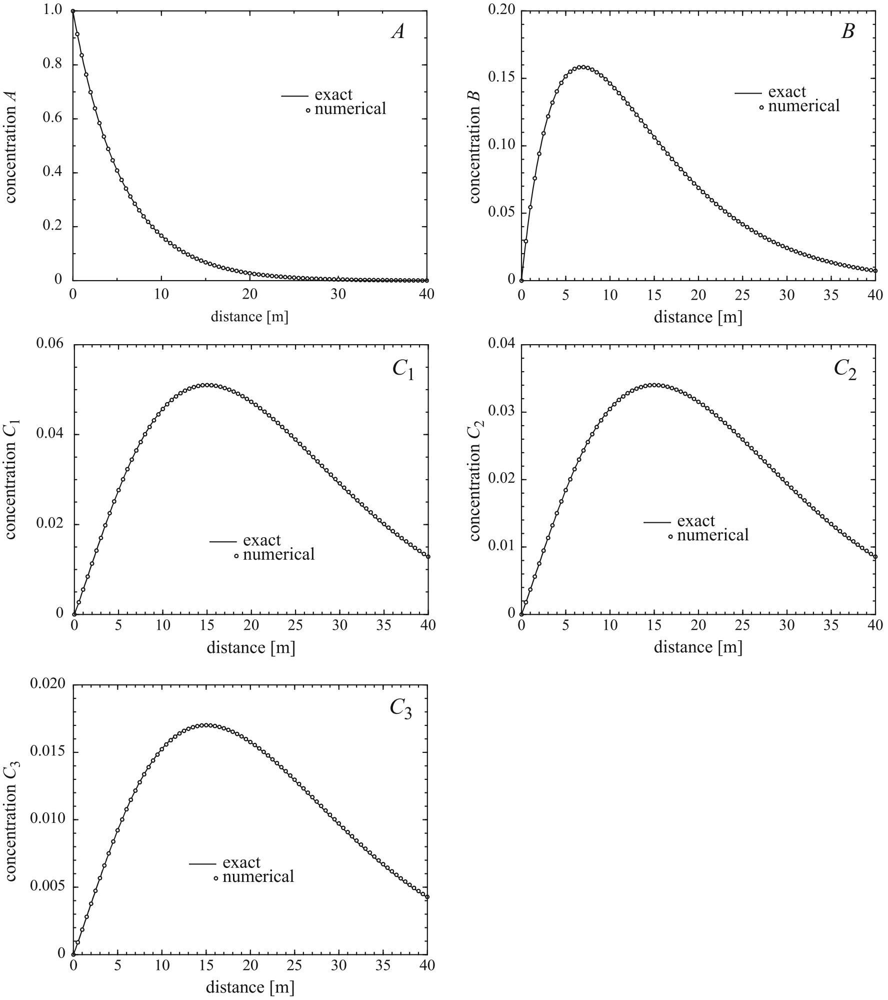


Fig. 12.8 Concentration profiles for the five species across the column of length L after $t = 40$ days of serial-parallel reactive transport: comparison of Sun et al.'s exact (analytical) solution to FEFLOW's numerical results

First-order reaction rates for the sequential reaction kinetics $C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4$ are given as follows

$$\begin{aligned} R_1 &= -k_1\,C_1 \\ R_2 &= k_1\,C_1 - k_2\,C_2 \\ R_3 &= k_2\,C_2 - k_3\,C_3 \\ R_4 &= k_3\,C_3 - k_4\,C_4 \end{aligned} \tag{12.70}$$

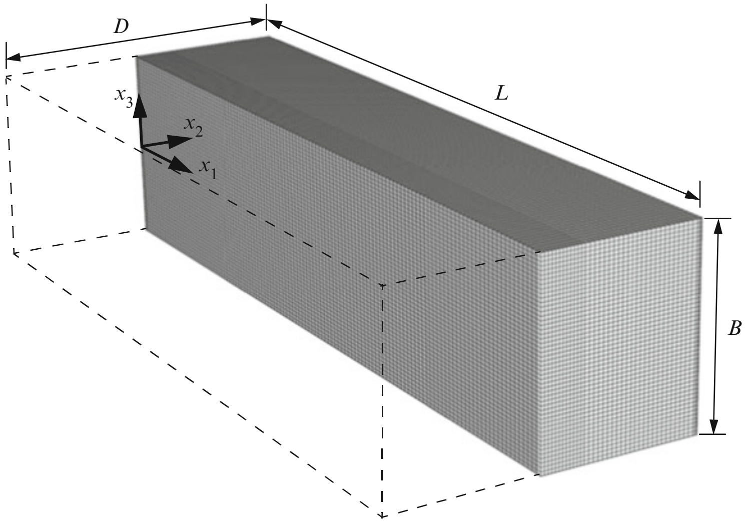


Fig. 12.9 Discretized 3D aquifer for Sun et al.'s problem. Structured mesh consists of 120×37×50 brick elements for the symmetric half-domain $L \times \frac{D}{2} \times B$

where k_k $(k = 1, 2, 3, 4)$ are rate constants, which are listed together with the remaining parameters in Table 12.5. All species are considered mobile in the liquid phase; porosity ε does not play a role. Unspecified BC's represent boundaries, at which natural BC's are imposed.

The obtained results are shown in Figs. 12.10 and 12.11. The concentration contours reveal differences between the analytical and numerical solutions. In the finite-element analysis the aquifer is finite and natural Neumann-type BC's (zero concentration gradients) are applied at the outer border faces of the discretized 3D domain. Unlike, in the analytical solution the aquifer domain is considered semi-infinite. Furthermore, Sun et al. [503] used an alternative dispersion model, where different transverse dispersivities in the horizontal and vertical directions are applied. In the FEFLOW simulations the Bear-Scheidegger dispersion model (12.3) is preferred with only one transverse dispersion parameter (Table 12.5).

12.5.3.3 Rate-Limited Desorption and Decay: Comparison to Fry et al.'s Analytical Solution

Fry et al. [179] studied rate-limited desorption and 1st-order decay on the feasibility of in situ bioremediation of contaminated groundwater by using analytical solutions. The conceptual model is shown in Fig. 12.12. A remedial pump-and-treat scheme is considered assuming conditions of 1D, steady-state groundwater flow through a homogeneous and isotropic aquifer. The modeled portion of the aquifer is bounded by injection and extraction wells (see control volume drawn in Fig. 12.12).

Table 12.5 Problem parameters used for the 3D multispecies mass transport with 1st-order degradation reaction kinetics

Quantity	Symbol	Value	Unit
Domain measure (length; half-with; thickness)[a]	$L; \frac{D}{2}; B$	$100; \frac{41}{2}; 25$	m
Longitudinal dispersivity	β_L	1.5	m
Transverse dispersivity	β_T	0.3	m
Pore velocity	v	0.2	$\mathrm{m\,d^{-1}}$
Molecular diffusion ($k = 1, 2, 3, 4$)	D_k	0	$\mathrm{m^2\,s^{-1}}$
Rate constant of species C_1	k_1	0.05	$\mathrm{d^{-1}}$
Rate constant of species C_2	k_2	0.02	$\mathrm{d^{-1}}$
Rate constant of species C_3	k_3	0.01	$\mathrm{d^{-1}}$
Rate constant of species C_4	k_4	0.005	$\mathrm{d^{-1}}$
IC's and BC's			
Initial condition (IC) of C_k ($k = 1, 2, 3, 4$)	$C_{k,0}$	0	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_1 at ($x_1 = 0, 0 \le x_2 \le 5.5\,\mathrm{m}, -2.5\,\mathrm{m} \le x_3 \le 2.5\,\mathrm{m}$)	C_{1D}	1	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_1 at ($x_1 = 0, x_2 > 5.5\,\mathrm{m}, x_3 < -2.5\,\mathrm{m}, x_3 > 2.5\,\mathrm{m}$)	C_{1D}	0	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_k ($k = 2, 3, 4$) at ($x_1 = 0, 0 \le x_2 \le 20.5\,\mathrm{m}, -12.5\,\mathrm{m} \le x_3 \le 12.5\,\mathrm{m}$)	C_{kD}	0	$\mathrm{mg\,l^{-1}}$
FEM			
3D mesh of 120 × 37 × 50 brick elements (Fig. 12.9), GFEM and AB/TR			
Initial time step size	Δt_0	10^{-7}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	400	d

[a] Measures and origin of coordinate system (x_1, x_2, x_3) defined in Fig. 12.9

The study concerns a method of restoration of a contaminated aquifer domain, where organic compounds are degraded by indigenous or introduced microorganisms. Degradation of the contaminant is represented by a 1st-order decay, where the rate of degradation is a function of the contaminant concentration in the aqueous (liquid) phase. Desorption is described using 1st-order kinetics, where the rate of mass transfer of contaminant from the solid phase to the aqueous phase depends on the concentration gradient between the two phases and a single rate coefficient. The following 1D two-species transport equations are considered (written in the present notation), which is a *one-site kinetic model* [351, 542] with linear kinetic sorption and decay in the aqueous phase:

$$\begin{aligned} \varepsilon\frac{\partial C}{\partial t} + \varepsilon_s\frac{\partial S}{\partial t} + q\frac{\partial C}{\partial x} - \mathcal{D}\frac{\partial^2 C}{\partial x^2} &= -\varepsilon\vartheta C \\ \varepsilon_s\frac{\partial S}{\partial t} &= \varepsilon_s\alpha(\rho^s K^d C - S) \end{aligned} \tag{12.71}$$

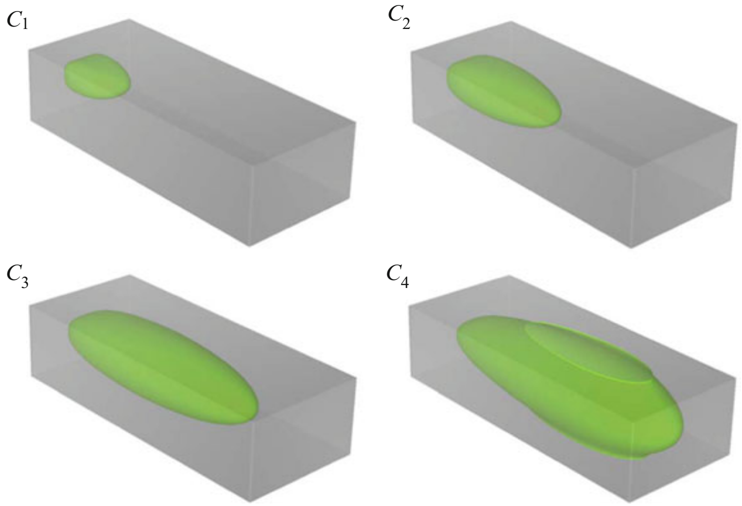


Fig. 12.10 FEFLOW results of the 3D 0.01 isosurface concentration for the four species C_1, C_2, C_3 and C_4 after $t = 400$ days

or

$$\begin{aligned} \varepsilon\frac{\partial C}{\partial t} + q\frac{\partial C}{\partial x} - \mathcal{D}\frac{\partial^2 C}{\partial x^2} &= R_C \\ \varepsilon_s\frac{\partial S}{\partial t} &= R_S \end{aligned} \tag{12.72}$$

with

$$\begin{aligned} R_C &= -(\varepsilon\vartheta + \varepsilon_s\rho^s\alpha K^d)C + \varepsilon_s\alpha S \\ R_S &= \varepsilon_s\alpha(\rho^s K^d C - S) \end{aligned} \tag{12.73}$$

where C is the aqueous concentration (at liquid phase), S is the sorbed concentration (at solid phase), $\mathcal{D} = \varepsilon D + \beta_L q$ is the hydrodynamic dispersion coefficient, $\varepsilon_s = 1 - \varepsilon$ is the solid volume fraction, ρ^s is the solid density, K^d is the distribution coefficient (cf. Table 3.8) and α is the 1st-order desorption rate constant.

The aquifer is initially contaminated and concentrations C and S are uniform throughout the control volume. Furthermore, the sorbed and aqueous phases are initially in linear equilibrium as described with the distribution coefficient K^d. These IC's are stated as

$$\begin{aligned} C(x,0) &= C_0 \qquad && (0 \le x \le L) \\ S(x,0) &= S_0 \qquad && (0 \le x \le L) \\ S_0 &= \rho^s K^d C_0 \end{aligned} \tag{12.74}$$

where C_0 is the aqueous concentration at $t = 0$, S_0 is the sorbed concentration at $t = 0$ and L is the length of the control volume (Fig. 12.12).

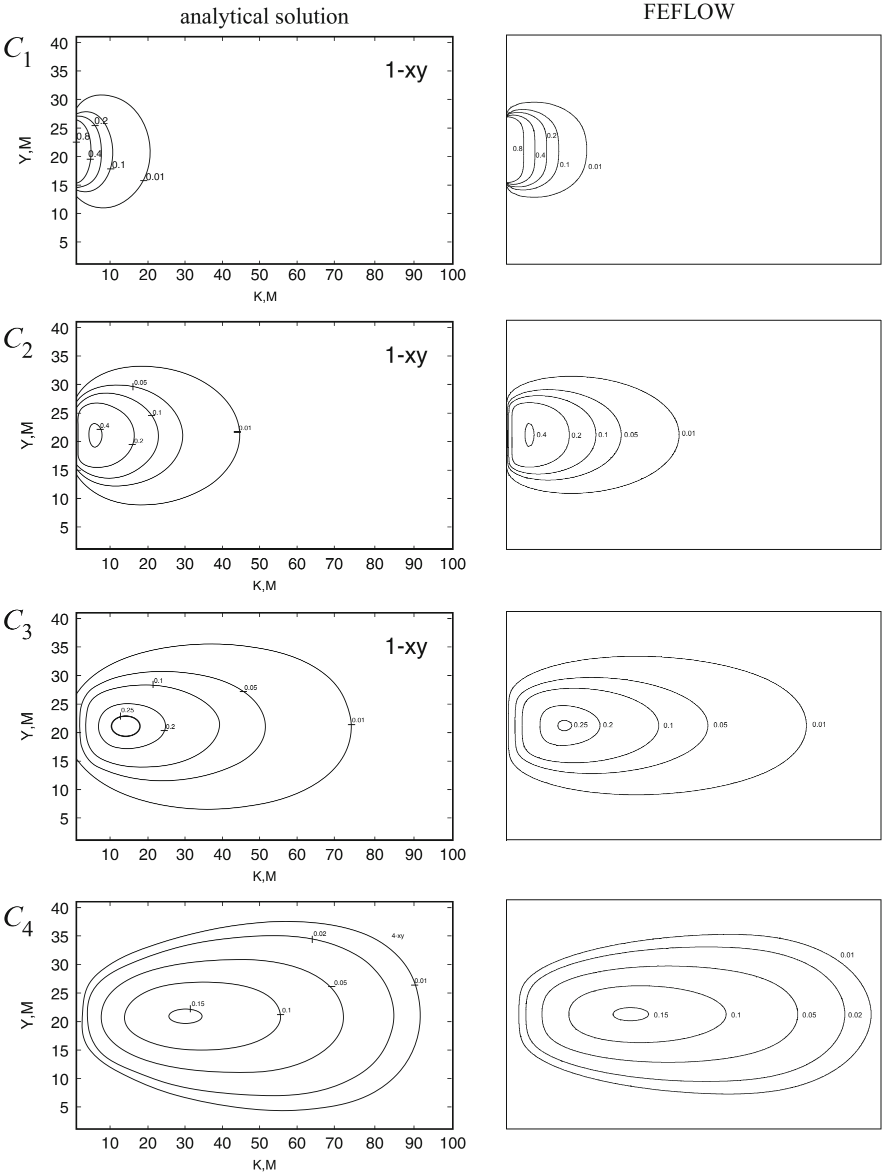


Fig. 12.11 Comparison of Sun et al.'s analytical solution [503] (*left*) with FEFLOW results (*right*): concentration contours of the four species C_1, C_2, C_3 and C_4 in the $x_1 - x_2$−plane at $x_3 = 13\,\text{m}$ and $t = 400$ days

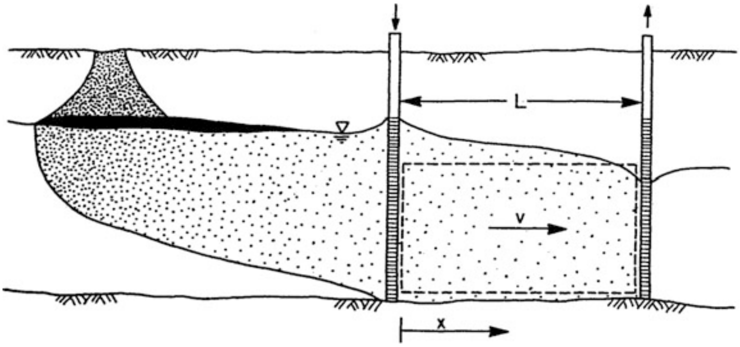


Fig. 12.12 Conceptual model of reacting contaminant transport in groundwater by Fry et al. [179]

At the control-volume inlet ($x = 0$) the contaminant flux due to advection and dispersion is zero at all times. At the control-volume exit ($x = L$) the concentrations are uniform with distance. Thus, the following BC's hold:

$$\begin{aligned} -\mathcal{D}\frac{\partial C}{\partial x}(0,t) + qC(0,t) &= 0 \qquad t > 0 \\ -\mathcal{D}\frac{\partial C}{\partial x}(L,t) &= 0 \qquad t > 0 \end{aligned} \tag{12.75}$$

A test case is considered for which the used parameters and conditions are listed in Table 12.6. Due to the BC's (12.75) the divergence form of the governing transport equations is used, which allows the input of the total (advective plus dispersive) mass flux at the boundary.

The reaction kinetics for the present problem is of a *degradation type*. We employ FEFLOW's reaction kinetics editor (see Sect. 5.5.4) to input the reaction rates (12.73), which are specified as follows (note that species ID 1 represents the aqueous species with concentration $C \equiv C_1$ and species ID 2 represents the sorbed species with concentration $S \equiv C_2$):

$$\begin{aligned} R_1 &= -(\text{Porosity}_1 \cdot \text{Rate}_1 + \text{Rate}_2 \cdot \text{Rb} \cdot \text{Kd}) \cdot C_1 + \text{Rate}_2 \cdot \text{SolidFrac}_2 \cdot C_2 \\ &\quad \text{Rb} = \text{SolidFrac}_2 \cdot 2.67 \quad \text{Kd} = 0.68 \\ R_2 &= \text{SolidFrac}_2 \cdot \text{Rate}_2 \cdot (\text{Rs} \cdot \text{Kd} \cdot C_1 - C_2) \\ &\quad \text{Rs} = 2.67 \quad \text{Kd} = 0.68 \end{aligned} \tag{12.76}$$

The parameters in (12.76) are related to the notation used in (12.73) as follows: $R_1 \equiv R_C$, $R_2 \equiv R_S$, $\text{Porosity}_1 \equiv \varepsilon$, $\text{Rate}_1 \equiv \vartheta$, $\text{Rate}_2 \equiv \alpha$, $\text{Rs} \equiv \rho^s$, $\text{Rb} \equiv \varepsilon_s \rho^s$, $\text{Kd} \equiv K^d$ and $\text{SolidFrac}_2 \equiv \varepsilon_s$. The species ID's are linked to the species names and phases as summarized in Table 12.7. The FEFLOW results for the problem are compared with the analytical solutions which are presented by Fry et al. [179].

Table 12.6 Parameters and conditions used for Fry et al.'s rate-limited desorption and decay transport problem in a column

Quantity	Symbol	Value	Unit
Column length	L	10	m
Constant flux	q	0.04	$\mathrm{m\,d^{-1}}$
Porosity	ε	0.4	1
Solid volume fraction	$\varepsilon_s = 1 - \varepsilon$	0.6	1
Distribution coefficient	K^d	0.68	$\mathrm{cm^3\,g^{-1}}$
Solid density	ρ^s	2.67	$\mathrm{g\,cm^{-3}}$
Decay rate	ϑ	0.1	$\mathrm{d^{-1}}$
Desorption rate constant	α	0.01	$\mathrm{d^{-1}}$
Molecular diffusion	D	0	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	1	m
Dispersion	$\mathcal{D} = \varepsilon D + \beta_L q$	$4.63 \cdot 10^{-7}$	$\mathrm{m^2\,s^{-1}}$
IC's and BC's			
Initial condition (IC) of C	C_0	1	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of S	S_0	1.816	$\mathrm{mg\,l^{-1}}$
Natural BC of C at $x = 0$	$q_{n_C} = -\mathcal{D}\nabla C \cdot \boldsymbol{n} + qC\vert_{x=0}$	0	$\mathrm{gm^{-2}\,d^{-1}}$
Natural BC of C at $x = L$	$q_{n_C} = -\mathcal{D}\nabla C \cdot \boldsymbol{n}$	0	$\mathrm{gm^{-2}\,d^{-1}}$
FEM			
Uniform mesh consisting of 300 linear elements, GFEM			
Space increment	Δx	0.03333	m
Initial time step size[a]	Δt_0	10^{-5}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	10^3	d

[a] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$ and maximum time step size $\Delta t_{\max} = 0.5$ day

Table 12.7 Species ID's used in Fry et al.'s 1D problem

ID ($= k$)	Phase	Name
1	Liquid	C
2	Solid	S

As shown in Fig. 12.13 very good agreement with the analytical results is obtained. The FEFLOW simulation takes 2,039 time steps of variable length.

12.5.3.4 Two-Site Equilibrium/Kinetic Sorption with Degradation: Comparison to STANMOD Analytical Solutions

The two-site sorption concept presumes that sorption or exchange sites in soils can be classified into two fractions: one fraction (Type-1) on which sorption is assumed to be instantaneous, and another fraction (Type-2) on which sorption is considered to be time-dependent. The resulting two-site kinetic model interacts with a solid phase composed of such different constituents as soil minerals, organic matter and various

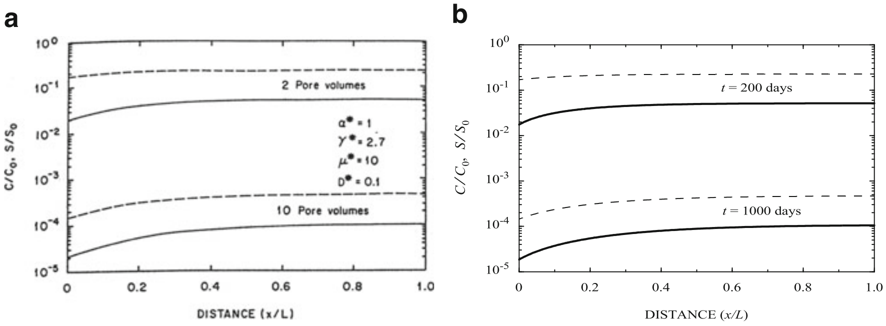


Fig. 12.13 Aqueous (C/C_0, *solid lines*) and sorbed (S/S_0, *dashed lines*) concentrations versus distance x at times $t = 200$ days and $t = 1{,}000$ days: (**a**) Fry et al.'s analytical solution [179] (pore volumes $= tq/(\varepsilon L)$) and (**b**) FEFLOW results

oxides. Studies in transport of pesticides indicate that the two-site kinetic model may well be suitable [542].

The derivation proceeds in the same fashion as for the one-site sorption model of the preceding Sect. 12.5.3.3. We introduce two different sorbed concentrations S_1 and S_2, where the first one is for Type-1 at equilibrium sites and the second one is for Type-2 at kinetic sites. Because Type-1 sites are always at equilibrium, sorption onto these sites is given by an adsorption function similar to (5.64), viz.,

$$S_1 = f\varphi C \tag{12.77}$$

where C is the aqueous concentration at liquid phase, f is the fraction of exchange sites assumed to be at equilibrium and φ is a sorption function. The kinetic part S_2 is subjected to a kinetic relationship in a form

$$S_2 \rightarrow (1 - f)\varphi C \tag{12.78}$$

By using the equilibrium sorption (12.77) the Type-1 concentration S_1 can be eliminated (expressed by C) from the 3-species basic equations and only 2 species (namely C and S_2) have to be solved. Assuming a linear degradation for all species C, S_1 and S_2, as well as a Henry-type sorption for S_1, we found the following 2-species model equations for a *two-site kinetic sorption* [517,542] with degradation written in the present notation:

$$\begin{aligned} \varepsilon s \Re \frac{\partial C}{\partial t} + \boldsymbol{q} \cdot \nabla C - \nabla \cdot (\boldsymbol{D} \cdot \nabla C) &= R_C \\ \varepsilon_s \frac{\partial S_2}{\partial t} &= R_S \end{aligned} \tag{12.79}$$

with

$$\begin{aligned} R_C &= -\big[\alpha\varepsilon_s \tfrac{(1-f)}{f}\kappa + \varepsilon_s\kappa\vartheta_{S_1} + \varepsilon s\vartheta_C\big]C + \alpha\varepsilon_s S_2 \\ R_S &= \alpha\varepsilon_s \tfrac{(1-f)}{f}\kappa C - \varepsilon_s(\alpha + \vartheta_{S_2})S_2 \end{aligned} \tag{12.80}$$

and (cf. Table 3.8)

$$\begin{aligned} \Re &= 1 + \big(\tfrac{1-\varepsilon}{\varepsilon}\big)\kappa \\ \kappa &= f\rho^s K^d \end{aligned} \tag{12.81}$$

where C is the aqueous concentration (at liquid phase), S_1 is the Type-1 sorbed concentration (at solid phase), S_2 is the Type-2 sorbed concentration (at solid phase), $\varepsilon_s = 1 - \varepsilon$ is the solid volume fraction, ρ^s is the solid density, K^d is the distribution coefficient (cf. Table 3.8), f is the fraction of exchange sites, α is the 1st-order kinetic rate coefficient, κ is the Henry adsorption coefficient (cf. Table 3.8), ϑ_{S_1} is the decay coefficient of sorbed species S_1, ϑ_{S_2} is the decay coefficient of sorbed species S_2 and ϑ_C is the decay coefficient of diluted species C. Note that the two-site adsorption model (12.80) reduces to the one-site fully kinetic adsorption model comparable to (12.73) if $f \to 0$, where the $(1-f)\kappa/f$ terms in (12.80) have to be replaced by $(1-f)\rho^s K^d$.

We solve the above two-site kinetic sorption equations for a 1D domain (column) of length L, for which analytical solutions are available [517,527,542]. To compare to analytical solutions the following dimensionless parameters are to be defined:

$$\beta = \frac{\Re}{\Re^\star}, \quad \Re^\star = 1 + \Big(\frac{1-\varepsilon}{\varepsilon}\Big)\frac{\kappa}{f}, \quad \omega = \alpha(1-\beta)\Re^\star\frac{L}{v}, \quad \mathrm{Pe} = \frac{qL}{\mathcal{D}} \tag{12.82}$$

where $q = \|\boldsymbol{q}\|$ is the constant 1D flux, $v = q/(s\varepsilon)$ is the constant 1D pore velocity and $\mathcal{D} = \|\boldsymbol{D}\|$ is the dispersion coefficient. With given parameters $\Re^\star$, β and ω, the model parameters κ, f and α can be specified as

$$\kappa = \frac{\varepsilon\beta(\Re^\star - 1) - (1-\beta)\varepsilon}{1-\varepsilon}, \quad f = \frac{\kappa(1-\varepsilon)}{(\Re^\star - 1)\varepsilon}, \quad \alpha = \frac{\omega}{(1-\beta)\Re^\star\frac{v}{L}} \tag{12.83}$$

Note that α is only defined if $\beta < 1$. Equations (12.79) with (12.80) are solved for an initially solute-free column subject to a pulse-type input BC. The IC's and BC's are stated as

$$\begin{aligned} C(x,0) &= 0 \qquad (0 \le x \le L) \\ S_2(x,0) &= 0 \qquad (0 \le x \le L) \end{aligned} \tag{12.84}$$

and

$$\begin{aligned} -\mathcal{D}\frac{\partial C}{\partial x}(0,t) + qC(0,t) &= \begin{cases} -qC_o & 0 < t \le t_o \\ 0 & t > t_o \end{cases} \\ -\mathcal{D}\frac{\partial C}{\partial x}(L,t) &= 0 \qquad\qquad t > 0 \end{aligned} \tag{12.85}$$

where C_o is the input concentration and t_o is the time duration of the applied solute pulse.

We consider the 1D column for a steady-state flow (q = const) and saturated conditions ($s = 1$). Furthermore, we assume that all decay coefficients are the same, i.e., $\vartheta = \vartheta_C = \vartheta_{S_1} = \vartheta_{S_2}$. Accordingly, a dimensionless decay parameter ξ is defined as

$$\xi = \frac{\vartheta L}{v} \tag{12.86}$$

The test case is considered for the dimensionless parameters with $\beta = 0.5$, $\Re^\star = 2.5$, $\omega = 0.5$, $\mathrm{Pe} = 4.7$ and $\xi \in (0; 0.1; 0.3; 0.6; 1.0)$. In accordance with (12.83) the complete dataset used for the numerical simulation is listed in Table 12.8. Due to the BC's (12.85) the divergence form of the governing transport equations is used, which allows the input of the total (advective plus dispersive) mass flux at the boundary.

The reaction kinetics for the two-site kinetic transport problem is of a *degradation type*. We again prefer FEFLOW's reaction kinetics editor (see Sect. 5.5.4) to input the reaction rates (12.80), which are specified as follows (note that species ID 1 represents the aqueous species with concentration $C \equiv C_1$ and species ID 2 represents the sorbed species with concentration $S_2 \equiv C_2$):

$$\begin{aligned}
R_1 &= -(\mathrm{Rate}_2 \cdot \mathrm{SolidFrac}_2 \cdot g \cdot K + \mathrm{SolidFrac}_2 \cdot K \cdot \mathrm{Rate}_1 + \\
&\quad \mathrm{Porosity}_1 \cdot \mathrm{Rate}_1) \cdot C_1 + \mathrm{Rate}_2 \cdot \mathrm{SolidFrac}_2 \cdot C_2 \\
&\quad f = 0.16667 \quad g = (1-f)/f \quad K = \mathrm{Sorption}_1 \\
R_2 &= \mathrm{Rate}_2 \cdot \mathrm{SolidFrac}_2 \cdot g \cdot K \cdot C_1 - \mathrm{SolidFrac}_2 \cdot (\mathrm{Rate}_2 + \mathrm{Rate}_1) \cdot C_2 \\
&\quad f = 0.16667 \quad g = (1-f)/f \quad K = \mathrm{Sorption}_1
\end{aligned} \tag{12.87}$$

The parameters in (12.87) are related to the notation used in (12.80) as follows: $R_1 \equiv R_C$, $R_2 \equiv R_S$, $\mathrm{Porosity}_1 \equiv \varepsilon$, $\mathrm{Rate}_1 \equiv \vartheta$, $\mathrm{Rate}_2 \equiv \alpha$, $\mathrm{Sorption}_1 \equiv \kappa$ and $\mathrm{SolidFrac}_2 \equiv \varepsilon_s$. The species ID's are linked to the species names and phases as summarized in Table 12.9.

The FEFLOW results for the problem are compared with the analytical solutions which are evaluated by using the STANMOD package [527]. We simulate the breakthrough characteristics for C and S_2 measured at the effluent boundary at $x = L$ for different decay parameters ξ (12.86). The plots are related to dimensionless aqueous and sorbed concentrations and , respectively, defined as

$$\hat{C} = \frac{C}{C_o}, \qquad \hat{S}_2 = \frac{S_2}{(\frac{1-f}{f})\kappa C_o} \tag{12.88}$$

Figure 12.14 reveals a good agreement with the analytical solutions. The required number of adaptive AB/TR time steps is about 230.

Table 12.8 Parameters and conditions used for the two-site kinetic transport problem in a column

Quantity	Symbol	Value	Unit
Column length	L	10	m
Constant flux	q	0.04	$\mathrm{m\,d^{-1}}$
Porosity	ε	0.4	1
Pore velocity	$v = \frac{q}{\varepsilon}$	0.1	$\mathrm{m\,d^{-1}}$
Solid volume fraction	$\varepsilon_s = 1 - \varepsilon$	0.6	1
Henry coefficient	κ	0.16667	1
Fraction of exchange site	f	0.16667	1
Kinetic rate coefficient	α	0.004	$\mathrm{d^{-1}}$
Decay rate coefficient	ϑ	$\begin{cases} 0 \\ 0.001 \\ 0.003 \\ 0.006 \\ 0.01 \end{cases}$	$\mathrm{d^{-1}}$
Molecular diffusion	D	0	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	2.128	m
Dispersion	$\mathcal{D} = \varepsilon D + \beta_L q$	$9.85 \cdot 10^{-7}$	$\mathrm{m^2\,s^{-1}}$
IC's and BC's			
Initial condition (IC) of C	C_0	0	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of S_2	S_0	0	$\mathrm{mg\,l^{-1}}$
Input concentration of C	C_o	1	$\mathrm{mg\,l^{-1}}$
Pulse duration	t_o	300	d
Neumann-type BC of C at $x = 0$	$q_{n_C} = -\mathcal{D}\nabla C \cdot \boldsymbol{n} + qC\vert_{x=0}$	$\begin{cases} -qC_o & 0 < t \le t_o \\ 0 & t > t_o \end{cases}$	$\mathrm{gm^{-2}\,d^{-1}}$
Natural BC of C at $x = L$	$q_{n_C} = -\mathcal{D}\nabla C \cdot \boldsymbol{n}$	0	$\mathrm{gm^{-2}\,d^{-1}}$
FEM			
Uniform mesh consisting of 300 linear elements, GFEM			
Space increment	Δx	0.03333	m
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	800	d

Table 12.9 Species ID's used for the two-site kinetic transport problem

ID ($= k$)	Phase	Name
1	Liquid	C
2	Solid	S_2

12.5.4 Multispecies Mass Transport of Sequential and Nonsequential Chlorinated Solvents Degradation Under Variable Aerobic-Anaerobic Conditions

In contrast to nonsequential (aerobic) degradation of chlorinated solvents, sequential dehalogenation is performed by anaerobic bacteria that cannot work under aerobic

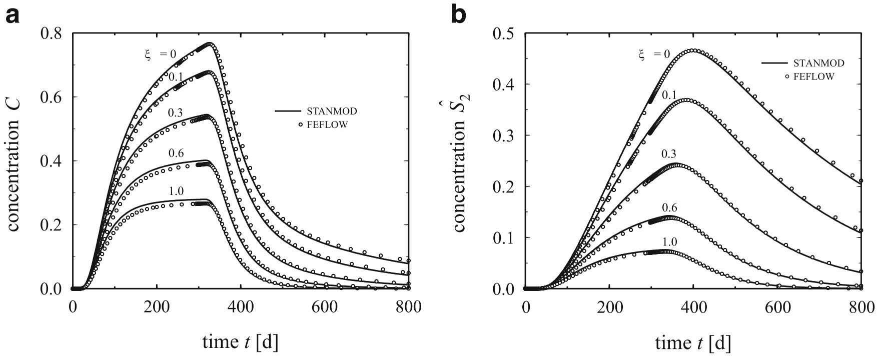


Fig. 12.14 FEFLOW results versus STANMOD solutions [527] for effluent breakthrough history at $x = L$ of (**a**) aqueous $\hat{C}$ and (**b**) sorbed $\hat{S}_2$ concentrations for different decay parameters ξ at $\mathrm{Pe} = 4.7$, $\beta = 0.5$, $\omega = 0.5$, $\Re = 1.25$ and $t_o = 300$ days

conditions.[9] Both mechanisms can occur in the same contaminant plume depending on oxygen and nitrate concentrations. Monitoring the chloride released during the dehalogenation can be useful to locate the areas where dehalogenation occurs and to estimate degradation rates. This example simulation issues from a benchmark within the MACAOH (Modélisation, Atténuation, Charactérisation dans les Aquiféres des composés Organo-Halogénés) project [4] of the French Environment and Energy Management Agency (ADEME) with various university and private partners. The project focuses on chlorinated solvents, specifically PCE (perchloroethylene), TCE (trichloroethylene), DCE (cis- and trans-1,2-dichloroethylene) and VC (vinyl chloride). The aim of the benchmark was to evaluate the state of the art in the numerical simulation of the natural degradation of chlorinated solvents in aquifers in France.

A mixture of PCE and TCE is injected continuously in an initially uncontaminated 1D domain containing dissolved oxygen, nitrate and chloride. The initial aerobic conditions do not allow the degradation of PCE by anaerobic bacteria, but a slow complete mineralization of TCE is considered. Daughter products of TCE during the mineralization (H_2O and CO_2) are not simulated, except chloride ions. The oxygen concentration decreases as a consequence of the aerobic bacteria respiration. This behavior continues as long as the concentration of oxygen remains above a critical level. The reactions take place only in the liquid phase. Reaction and sorption with the solid phase are neglected.

Wiedemeier et al. [564] explained that anaerobic bacteria cannot work at oxygen concentrations greater than $0.5\,\mathrm{mg\,l^{-1}}$. When no more oxygen remains in water, aerobic bacteria use nitrate. After Wiedemeier et al. [564], anaerobic bacteria can

[9] FEFLOW results described in this section were obtained by D. Etcheverry† and Y. Rossier (France).

start the sequential reductive dechlorination of chlorinated solvents under nitrate concentrations smaller than 1 mg l^{-1}.

In the MACAOH benchmark, it was assumed that the sequential degradation starts in the presence of nitrate as soon as the oxygen concentration reaches zero. Thus, once oxygen reaches sufficiently low concentration anywhere in the domain, the following reductive sequential degradation of chlorinated solvents starts

$$\mathrm{PCE} \overset{H^+,2e^-}{\longrightarrow} \mathrm{TCE} + \mathrm{Cl}^- \overset{H^+,2e^-}{\longrightarrow} \mathrm{DCE} + \mathrm{Cl}^- \overset{H^+,2e^-}{\longrightarrow} \mathrm{VC} + \mathrm{Cl}^- \tag{12.89}$$

The concentration of nitrate is supposed to decrease independently of other species, as soon as the oxygen concentration reaches zero, following a simple 1st-order degradation law.

The 1D equations of transport for the homogeneous reaction of nonretarded parent and daughter species can be written as

$$\begin{aligned} \varepsilon\frac{\partial C_k}{\partial t} - \mathcal{D}\frac{\partial^2 C_k}{\partial x^2} + q\frac{\partial C_k}{\partial x} &= \varepsilon\left[\delta_k k_k^{\mathrm{anae}} C_k - (1-\delta_k) k_k^{\mathrm{ae}} C_k\right] \\ \varepsilon\frac{\partial C_j}{\partial t} - \mathcal{D}\frac{\partial^2 C_j}{\partial x^2} + q\frac{\partial C_j}{\partial x} &= \varepsilon\left[\delta_j k_j^{\mathrm{anae}} C_j - \delta_j \nu_{k,j} k_k^{\mathrm{anae}} C_k - (1-\delta_j) k_j^{\mathrm{ae}} C_j\right] \end{aligned} \tag{12.90}$$

where k denotes the parent species (PCE to DCE), j the daughter product (TCE to VC), C_k and C_j the concentration of species k and j, respectively, k_k^{ae}, k_j^{ae}, k_k^{anae} and k_j^{anae} the 1st-order decay constants of species k and j under aerobic and anaerobic conditions, respectively, $\nu_{k,j}$ the stoichiometric coefficient for the degradation of species k to produce species j, δ_k and δ_j functions equal to 0 for degradation in aerobic conditions and equal to 1 for degradation in anaerobic conditions.

The reaction rates appearing on RHS's of (12.90) are described as follows:

Chlorinated solvents:

Under aerobic conditions, there is no sequential degradation and the reaction rates $R^{\mathrm{ae}}_{\mathrm{PCE,TCE,DCE,VC}}$ of chlorinated solvents in aerobic conditions simplify to

$$\begin{aligned} R^{\mathrm{ae}}_{\mathrm{PCE}} &= 0 \\ R^{\mathrm{ae}}_{\mathrm{TCE}} &= -\varepsilon(k^{\mathrm{ae}}_{\mathrm{TCE}}\, C_{\mathrm{TCE}}) \\ R^{\mathrm{ae}}_{\mathrm{DCE}} &= -\varepsilon(k^{\mathrm{ae}}_{\mathrm{DCE}}\, C_{\mathrm{DCE}}) \\ R^{\mathrm{ae}}_{\mathrm{VC}} &= -\varepsilon(k^{\mathrm{ae}}_{\mathrm{VC}}\, C_{\mathrm{VC}}) \end{aligned} \tag{12.91}$$

Under anaerobic conditions, the sequential degradation from PCE to VC leads to reaction rates $R^{\mathrm{anae}}_{\mathrm{PCE,TCE,DCE,VC}}$ defined as follows

Table 12.10 1st-order decay rates and stoichiometric coefficients used for the MACAOH benchmark example

Species k	k_k^{anae} (d^{-1})	k_k^{ae} (d^{-1})	$\nu_{k,j}$ (1)
PCE	0.03	0	–
TCE	0.09	0.009	$\nu_{\text{PCE,TCE}} = 0.792$
DCE	0.009	0.15	$\nu_{\text{TCE,DCE}} = 0.738$
VC	0	0.24	$\nu_{\text{DCE,VC}} = 0.644$
O_2	0	0	–
$NO3^-$	0.1	0	–
Cl^-	0	0	–

$$\begin{aligned}
R_{\text{PCE}}^{\text{anae}} &= -\varepsilon(k_{\text{PCE}}^{\text{anae}} C_{\text{PCE}})\\
R_{\text{TCE}}^{\text{anae}} &= -\varepsilon(k_{\text{TCE}}^{\text{anae}} C_{\text{TCE}} - \nu_{\text{PCE,TCE}}\, k_{\text{PCE}}^{\text{anae}}\, C_{\text{PCE}})\\
R_{\text{DCE}}^{\text{anae}} &= -\varepsilon(k_{\text{DCE}}^{\text{anae}} C_{\text{DCE}} - \nu_{\text{TCE,DCE}}\, k_{\text{TCE}}^{\text{anae}}\, C_{\text{TCE}})\\
R_{\text{VC}}^{\text{anae}} &= -\varepsilon(k_{\text{VC}}^{\text{anae}} C_{\text{VC}} - \nu_{\text{DCE,VC}}\, k_{\text{DCE}}^{\text{anae}}\, C_{\text{DCE}})
\end{aligned} \tag{12.92}$$

Except for PCE that has no parent species, all reaction rates are made of an independent degradation term and of a production term dependent on the degradation of the parent species.

Oxygen:

Aerobic bacteria do not use oxygen in definite proportions during their respiration. The oxygen consumption was arbitrarily defined in the benchmark as follows

$$\frac{\partial C_{O_2}}{\partial t} = 4.5\frac{\partial C_{\text{TCE}}}{\partial t} + 4\frac{\partial C_{\text{DCE}}}{\partial t} + 3.5\frac{\partial C_{\text{VC}}}{\partial t} \tag{12.93}$$

From the conceptual model, the terms in DCE and VC are superfluous because they are not present in the system at the initial state and because TCE does not degrade into those compounds in aerobic conditions. Thus in this example simulation there cannot be DCE or VC under aerobic conditions and it follows from (12.93) that

$$R_{O_2}^{\text{ae}} = 4.5R_{\text{TCE}}^{\text{ae}} + 4R_{\text{DCE}}^{\text{ae}} + 3.5R_{\text{VC}}^{\text{ae}} \tag{12.94}$$

By definition there is no oxygen in anaerobic conditions, so that

$$R_{O_2}^{\text{anae}} = 0 \tag{12.95}$$

Chloride:

Chloride is released into the groundwater during the dehalogenation of chlorinated solvents. The relation

$$\frac{\partial C_{\text{Cl}^-}}{\partial t} = -1.068\frac{\partial C_{\text{TCE}}}{\partial t} - 0.712\frac{\partial C_{\text{DCE}}}{\partial t} - 0.552\frac{\partial C_{\text{VC}}}{\partial t} \tag{12.96}$$

Table 12.11 Parameters and conditions used for the MACAOH benchmark example

Quantity	Symbol	Value	Unit
Column length	L	250	m
Constant flux	q	0.4	$\mathrm{m\,d^{-1}}$
Porosity	ε	0.4	1
Pore velocity	$v = \frac{q}{\varepsilon}$	1	$\mathrm{m\,d^{-1}}$
Decay rates and stoichiometric coefficients of species: $k =$ PCE, TCE, DCE, VC, O_2, $NO3^-$, Cl^- *are listed in Table 12.10*			
Longitudinal dispersivity	β_L	1	m
Retardation factor (no adsorption)	$\mathfrak{R} = \acute{\mathfrak{R}}$	1	1
Molecular diffusion	D	$1 \cdot 10^{-9}$	$\mathrm{m^2\,s^{-1}}$
Dispersion	$\mathcal{D} = \varepsilon D + \beta_L q$	$4.63 \cdot 10^{-6}$	$\mathrm{m^2\,s^{-1}}$
IC's and BC's			
Initial condition (IC) of C_{PCE}	$C_{\mathrm{PCE},0}$	0	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of C_{TCE}	$C_{\mathrm{TCE},0}$	0	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of C_{DCE}	$C_{\mathrm{DCE},0}$	0	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of C_{VC}	$C_{\mathrm{VC},0}$	0	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of $C_{\mathrm{O_2}}$	$C_{\mathrm{O_2},0}$	10	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of $C_{\mathrm{NO3^-}}$	$C_{\mathrm{NO3^-},0}$	20	$\mathrm{mg\,l^{-1}}$
Initial condition (IC) of $C_{\mathrm{Cl^-}}$	$C_{\mathrm{Cl^-},0}$	15	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_{PCE} at $x = 0$	$C_{\mathrm{PCE}\,D}$	3	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_{TCE} at $x = 0$	$C_{\mathrm{TCE}\,D}$	5	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_{DCE} at $x = 0$	$C_{\mathrm{DCE}\,D}$	0	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species C_{VC} at $x = 0$	$C_{\mathrm{VC}\,D}$	0	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species $C_{\mathrm{O_2}}$ at $x = 0$	$C_{\mathrm{O_2}\,D}$	10	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species $C_{\mathrm{NO3^-}}$ at $x = 0$	$C_{\mathrm{NO3^-}\,D}$	20	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC of species $C_{\mathrm{Cl^-}}$ at $x = 0$	$C_{\mathrm{Cl^-}\,D}$	15	$\mathrm{mg\,l^{-1}}$
Natural BC of all species k at $x = L$	$q_{n_{kC}} = -\mathcal{D}\nabla C_k \cdot \boldsymbol{n}$	0	$\mathrm{gm^{-2}\,d^{-1}}$
FEM			
Uniform mesh consisting of 250 linear elements, GFEM			
Space increment	Δx	1	m
Initial time step size[a]	Δt_0	10^{-3}	d
Maximum error tolerance (FE/BE)	ϵ	10^{-4}	1
Simulation time period	t_{end}	365	d

[a] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 1.1$ and maximum time step size $\Delta t_{\max} = 0.5$ day

Table 12.12 Reaction rates as defined in FEFLOW's reaction kinetics editor (cf. Sect. 5.5.4)

Reaction rate R_k ($k = \mathrm{PCE, TCE, DCE, VC, O_2, NO3^-, Cl^-}$)	Comment
$R_{\mathrm{PCE}} = \begin{cases} -\varepsilon(k^{\mathrm{anae}}_{\mathrm{PCE}} C_{\mathrm{PCE}}) & \text{if } C_{\mathrm{O_2}} < 0.1 \\ 0 & \text{otherwise} \end{cases}$	$k^{\mathrm{ae}}_{\mathrm{PCE}} = 0$
$R_{\mathrm{TCE}} = \begin{cases} -\varepsilon(k^{\mathrm{anae}}_{\mathrm{TCE}} C_{\mathrm{TCE}} - \nu_{\mathrm{PCE,TCE}} k^{\mathrm{anae}}_{\mathrm{PCE}} C_{\mathrm{PCE}}) & \text{if } C_{\mathrm{O_2}} < 0.1 \\ -\varepsilon(k^{\mathrm{ae}}_{\mathrm{TCE}} C_{\mathrm{TCE}}) & \text{otherwise} \end{cases}$	
$R_{\mathrm{DCE}} = \begin{cases} -\varepsilon(k^{\mathrm{anae}}_{\mathrm{DCE}} C_{\mathrm{DCE}} - \nu_{\mathrm{TCE,DCE}} k^{\mathrm{anae}}_{\mathrm{TCE}} C_{\mathrm{TCE}}) & \text{if } C_{\mathrm{O_2}} < 0.1 \\ -\varepsilon(k^{\mathrm{ae}}_{\mathrm{DCE}} C_{\mathrm{DCE}}) & \text{otherwise} \end{cases}$	Degradation is supposed under aerobic conditions
$R_{\mathrm{VC}} = \begin{cases} -\varepsilon(0 - \nu_{\mathrm{DCE,VC}} k^{\mathrm{anae}}_{\mathrm{DCE}} C_{\mathrm{DCE}}) & \text{if } C_{\mathrm{O_2}} < 0.1 \\ -\varepsilon(k^{\mathrm{ae}}_{\mathrm{VC}} C_{\mathrm{VC}}) & \text{otherwise} \end{cases}$	No VC decay in anaerobic conditions ($k^{\mathrm{anae}}_{\mathrm{VC}} = 0$) but decay supposed under aerobic conditions
$R_{\mathrm{O_2}} = \begin{cases} -\varepsilon(4.5 k^{\mathrm{ae}}_{\mathrm{TCE}} C_{\mathrm{TCE}} + 4 k^{\mathrm{ae}}_{\mathrm{DCE}} C_{\mathrm{DCE}} + \\ \quad 3.5 k^{\mathrm{ae}}_{\mathrm{VC}} C_{\mathrm{VC}}) & \text{if } C_{\mathrm{O_2}} > 0.05 \\ 0 & \text{otherwise} \end{cases}$	
$R_{\mathrm{NO3^-}} = \begin{cases} -\varepsilon(k^{\mathrm{anae}}_{\mathrm{NO3^-}} C_{\mathrm{NO3^-}}) & \text{if } C_{\mathrm{O_2}} < 0.1 \\ 0 & \text{otherwise} \end{cases}$	Decay of nitrate starts under anaerobic conditions
$R_{\mathrm{Cl^-}} = \begin{cases} \varepsilon(0.208 k^{\mathrm{anae}}_{\mathrm{PCE}} C_{\mathrm{PCE}} + 0.262 k^{\mathrm{anae}}_{\mathrm{TCE}} C_{\mathrm{TCE}} + \\ \quad 0.356 k^{\mathrm{anae}}_{\mathrm{DCE}} C_{\mathrm{DCE}} + 0.552 k^{\mathrm{anae}}_{\mathrm{VC}} C_{\mathrm{VC}}) & \text{if } C_{\mathrm{O_2}} < 0.1 \\ \varepsilon(1.068 k^{\mathrm{ae}}_{\mathrm{TCE}} C_{\mathrm{TCE}} + 0.712 k^{\mathrm{ae}}_{\mathrm{DCE}} C_{\mathrm{DCE}} + \\ \quad 0.552 k^{\mathrm{ae}}_{\mathrm{VC}} C_{\mathrm{VC}}) & \text{otherwise} \end{cases}$	

is assumed under aerobic conditions, i.e., in terms of reaction rates

$$R^{\mathrm{ae}}_{\mathrm{Cl^-}} = -1.068 R^{\mathrm{ae}}_{\mathrm{TCE}} - 0.712 R^{\mathrm{ae}}_{\mathrm{DCE}} - 0.552 R^{\mathrm{ae}}_{\mathrm{VC}} \tag{12.97}$$

For reductive dechlorination, different dechlorination kinetics is assumed, viz.,

$$\frac{\partial C_{\mathrm{Cl^-}}}{\partial t} = -0.208 \frac{\partial C_{\mathrm{PCE}}}{\partial t} - 0.262 \frac{\partial C_{\mathrm{TCE}}}{\partial t} - 0.356 \frac{\partial C_{\mathrm{DCE}}}{\partial t} - 0.552 \frac{\partial C_{\mathrm{VC}}}{\partial t} \tag{12.98}$$

and

$$R^{\mathrm{anae}}_{\mathrm{Cl^-}} = -0.208 R^{\mathrm{anae}}_{\mathrm{PCE}} - 0.262 R^{\mathrm{anae}}_{\mathrm{TCE}} - 0.356 R^{\mathrm{anae}}_{\mathrm{DCE}} - 0.552 R^{\mathrm{anae}}_{\mathrm{VC}} \tag{12.99}$$

Nitrate:

Nitrate is supposed to degrade at a given independent rate if the oxygen concentration is zero, which leads to the following reaction rates

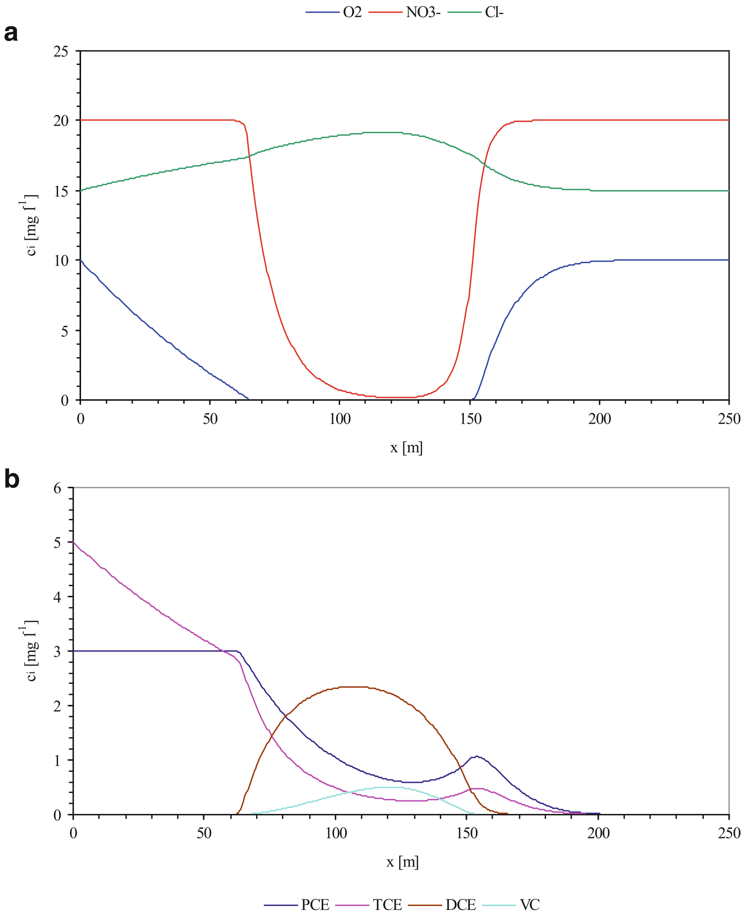


Fig. 12.15 Computed concentration profiles along the $x-$axis at 150 days for (**a**) oxygen, nitrate, chloride and (**b**) chlorinated solvents

$$\begin{aligned} R^{\text{ae}}_{\text{NO3}^-} &= 0 \\ R^{\text{anae}}_{\text{NO3}^-} &= -\varepsilon(k^{\text{anae}}_{\text{NO3}^-} C_{\text{NO3}^-}) \end{aligned} \tag{12.100}$$

1D steady flow and transient transport are supposed. The seven species simulated are PCE, TCE, DCE, VC, oxygen, nitrate and chloride. The aquifer length is 250 m, a constant Darcy flux q of 0.4 m d^{-1} is considered. For all species, porosity ε is 0.4, longitudinal dispersivity β_L is 1 m and retardation factor $\Re$ is 1. Steady Dirichlet-type species BC's of 3 mg l^{-1} for PCE, 5 mg l^{-1} for TCE, 0 mg l^{-1} for DCE and VC, 10 mg l^{-1} for oxygen, 20 mg l^{-1} for nitrate and 15 mg l^{-1} for chloride are applied at $x = 0$. IC's are assumed uniform, corresponding to the BC for oxygen, nitrate and chloride, and to zero for all chlorinated solvents. The 1st-order decay rates and

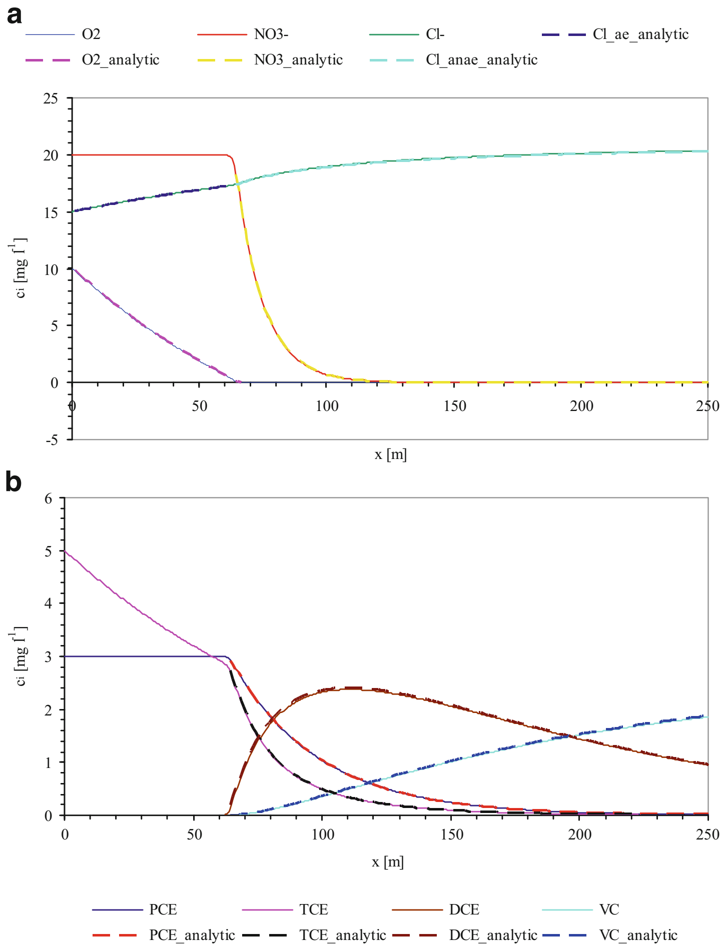


Fig. 12.16 Simulated versus analytical concentration profiles along the $x-$axis at steady state (365 days) for (**a**) oxygen, nitrate, chloride and (**b**) PCE, TCE, DCE, and VC. *Dashed lines* represent analytical solutions

stoichiometric coefficients are given in Table 12.10. Table 12.11 summarizes the simulation parameters and used conditions.

FEFLOW's versatile reaction kinetics editor (see Sect. 5.5.4) allows to easily define complex reaction rates. Particularly useful is the '*if otherwise*' construct to switch between aerobic and anaerobic behavior controlled via an oxygen concentration limit bounding the oxygen consumption below. It allows the combination of aerobic and anaerobic reaction rates in one composite reaction rate R_k for each species k. The reaction rates are entered in FEFLOW's reaction kinetics editor as listed in Table 12.12.

The simulation results are shown in Figs. 12.15 and 12.16 displaying the extent of the aerobic zone versus the anaerobic zone. At 150 days steady state is reached from $x = 0$ to $x = 65$ m for all species. Two separate aerobic zones appear from $x = 0$ to 65 m and from $x = 165$ m to the outlet. Anaerobic conditions are found between these two zones allowing the degradation of nitrate and the sequential degradation of PCE into TCE, DCE and VC. Under aerobic conditions the fast increase in chloride is a result of the complete mineralization kinetics of TCE. After 365 days the anaerobic zone extends from $x = 65$ m to the outlet. The sequential degradation of chlorinated solvents leads to an accumulation of VC. The analytical solutions are obtained for each species, for the aerobic domain first, then for the anaerobic domain, by applying the decoupling solution of Sun et al. [503] (cf. Sect. 12.5.3). No method was found to solve the problem analytically in transient state because the solution must include a switching term between aerobic and anaerobic rates of reactions as a function of the oxygen concentrations, which are varying in space and in time. On the other hand, steady state offers by definition a stable space and time limit between aerobic and anaerobic conditions. Thus, it is possible to solve first the aerobic domain, and to take the concentrations calculated at the end of the domain as BC's for the calculations in the anaerobic domain. To avoid the development of a solution for parallel reactions, the contributions of each chlorinated solvent to the production of chloride under anaerobic conditions are solved separately and added to the chloride concentration obtained at steady state at the end of the aerobic zone. As revealed in Fig. 12.16 the numerical results for all species are quasi identical to the analytical solutions.

Chapter 13
Heat Transport in Porous Media

13.1 Introduction

In this chapter we discuss the finite-element computation of heat (thermal energy) transport in porous media. Nonisothermal porous-medium processes can be found in many areas of application to natural and engineered systems, for instance exploitation of geothermal reservoirs as a viable and renewable source of energy, underground energy storage and recovery for heating and cooling purposes, waste disposal of heat-generating materials, chemical reactor engineering, insulation of buildings, material technology and many others. Modern industrial developments have expanded significantly the fields, where numerical simulation is required as a powerful tool to aid the design and operation of equipments. Of increasing interests are geothermal technologies in form of various heat exchanger systems in the underground to extract and/or to store thermal energy either in deep or even in shallow geologic formations by using open or closed geothermal systems, e.g., [25]. Here, geothermal heat pumps constitute a very attractive technology of a ground heat exchanger, equipped with a pipe network in which liquid (refrigerant) circulates in a closed loop, applied to shallow geothermal systems (usually not deeper than 250 m below the ground surface). Borehole heat exchanger (BHE) has shown a very suitable and cost-effective technology for both ground heat extraction and storage. However, its modeling requires specific concern due to the extreme slenderness of BHE's, where the pipes are in the order of 30 mm in diameter, the diameter of the borehole is in the order of 150 mm, the length of the borehole is in the order of 100 m and the extent of the model domain comprises hundreds or even thousands of meters.

The focus of the present chapter is in the formulation and numerical treatment of the governing porous-medium heat transport equations, including specific developments for BHE modeling. The solutions are associated with flow computations occurring in saturated or variably saturated porous media, under variable-density flow conditions or in combination with reactive mass transport which are subject of the preceding Chaps. 9–12, respectively.

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_13,

13.2 Basic Equations

13.2.1 3D, Vertical 2D and Axisymmetric Problems

The governing PDE's for 3D and vertical 2D (including axisymmetric) heat transport in porous media have been developed in Sect. 3.10.5 and summarized in Table 3.7. A major assumption is that the liquid and solid phases forming the porous medium are considered in a local thermodynamic equilibrium (3.241) leading to a unified system temperature T and a parallel behavior of thermal conductivities for the summed conservation equation of energy (first law of thermodynamics) written in its divergence form[1] as

$$\frac{\partial}{\partial t}\Big[\big(\varepsilon s\rho c+(1-\varepsilon)\rho^s c^s\big)(T-T_0)\Big]+\nabla\cdot(\rho c\boldsymbol{q}(T-T_0))-\nabla\cdot(\boldsymbol{\Lambda}\cdot\nabla T)=Q_T+Q_{Tw} \tag{13.1}$$

and its convective form as

$$\big(\varepsilon s\rho c+(1-\varepsilon)\rho^s c^s\big)\frac{\partial T}{\partial t}+\rho c\boldsymbol{q}\cdot\nabla T-\nabla\cdot(\boldsymbol{\Lambda}\cdot\nabla T)=Q_T+Q_{Tw}-\rho c(T-T_0)Q \tag{13.2}$$

associated with the constitutive relations

$$\begin{aligned}
\boldsymbol{\Lambda} &= \boldsymbol{\Lambda}_0+\boldsymbol{\Lambda}_0^s+\rho c\boldsymbol{D}_{\text{mech}} \\
\boldsymbol{\Lambda}_0 &= \varepsilon s\Lambda\boldsymbol{\delta} && \text{isotropic thermal conductivity of liquid} \\
\boldsymbol{\Lambda}_0^s &= \begin{cases}(1-\varepsilon)\Lambda^s\boldsymbol{\delta} \\ (1-\varepsilon)\Lambda^s\begin{pmatrix}1&0&0\\0&1&0\\0&0&\Xi_{\text{aniso}}^{\Lambda}\end{pmatrix}\end{cases} && \begin{aligned}&\text{isotropic thermal conductivity of solid}\\ &\text{anisotropic thermal conductivity of solid in 3D}\end{aligned} \\
\boldsymbol{D}_{\text{mech}} &= \beta_T\|\boldsymbol{q}\|\boldsymbol{\delta}+(\beta_L-\beta_T)\frac{\boldsymbol{q}\otimes\boldsymbol{q}}{\|\boldsymbol{q}\|} \\
Q_T &= \rho H^{\star}+\rho^s H_s^{\star}
\end{aligned} \tag{13.3}$$

where the source/sink term $H_e=\rho H+\rho^s H_s=Q_T+Q_{Tw}$ is suitably split into the supply term $Q_T=\rho H^{\star}+\rho^s H_s^{\star}$ and well-type SPC term Q_{Tw}. Similarly, the liquid sink/source term will be split by $Q=Q_h+Q_{hw}$. Furthermore, axis-parallel anisotropy for the solid thermal concuctivity $\boldsymbol{\Lambda}_0^s$ is optionally available in 3D according to (7.27) introducing the thermal anisotropy factor $\Xi_{\text{aniso}}^{\Lambda}$ (7.26).

The heat transport equation (13.1) or (13.2) has to be solved for the temperature T subject to a set of BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC (see Sect. 6.3.3), which is for the divergence form

[1] The divergence form (13.1) assumes that the specific heat capacities c and c^s are independent of T (cf. Sect. 3.9.1). Contrarily, the convective form (13.2) does not imply such an assumption.

$$
\begin{aligned}
T &= T_D && \text{on} \quad \Gamma_{D_T} \times t[t_0, \infty) \\
\big((T - T_0)\rho c \boldsymbol{q} - \boldsymbol{\Lambda} \cdot \nabla T\big) \cdot \boldsymbol{n} &= q_T^\dagger && \text{on} \quad \Gamma_{N_T} \times t[t_0, \infty) \\
\big((T - T_0)\rho c \boldsymbol{q} - \boldsymbol{\Lambda} \cdot \nabla T\big) \cdot \boldsymbol{n} &= -\Phi_T^\dagger (T_C - T) && \text{on} \quad \Gamma_{C_T} \times t[t_0, \infty) \\
Q_{Tw} &= -\textstyle\sum_w (T_w - T_0)\rho c Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty)
\end{aligned}
\tag{13.4}
$$

and for the convective form

$$
\begin{aligned}
T &= T_D && \text{on} \quad \Gamma_{D_T} \times t[t_0, \infty) \\
-(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} &= q_T && \text{on} \quad \Gamma_{N_T} \times t[t_0, \infty) \\
-(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} &= -\Phi_T (T_C - T) && \text{on} \quad \Gamma_{C_T} \times t[t_0, \infty) \\
Q_{Tw} &= -\textstyle\sum_w (T_w - T)\rho c Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) && \text{on} \quad \boldsymbol{x}_w \in \Omega \times t[t_0, \infty)
\end{aligned}
\tag{13.5}
$$

in combination with the IC of the form

$$
T(\boldsymbol{x}, t_0) = T_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\Omega}
\tag{13.6}
$$

where the total boundary is $\Gamma = \Gamma_{D_T} \cup \Gamma_{N_T} \cup \Gamma_{C_T}$. The normal heat fluxes on Γ_{N_T} and Γ_{C_T} differ between the divergence form and the convective form. As already discussed in Sects. 2.2.2 and 6.3.3 the divergence form imposes the total (advective plus conductive) boundary heat flux, while the convective form imposes a conductive heat flux on the boundary. However, the convective form can also be used to express a heat flux BC of an advective load by specifying the Cauchy-type BC in the form

$$
-(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} = \underbrace{-\Phi_T}_{\rho c \boldsymbol{q} \cdot \boldsymbol{n}} (\underbrace{T_C}_{\frac{q_T^\dagger}{\rho c \boldsymbol{q} \cdot \boldsymbol{n}} + T_0} - T)
\tag{13.7}
$$

to obtain

$$
\big((T - T_0)\rho c \boldsymbol{q} - \boldsymbol{\Lambda} \cdot \nabla T\big) \cdot \boldsymbol{n} = q_T^\dagger = \rho c \boldsymbol{q} \cdot \boldsymbol{n} (T_C - T_0)
\tag{13.8}
$$

for a given advective normal boundary flux $\boldsymbol{q} \cdot \boldsymbol{n}$ and a boundary temperature difference $T_C - T_0$, which is equivalent to a Neumann-type BC of the divergence form (cf. Sect. 6.3.3.3). Note further that OBC as discussed in Sect. 6.5.7 represents a special form of Neumann-type BC on $\Gamma_{N_O} \subset \Gamma_{N_T} \subset \Gamma$, which will be treated either as a natural Neumann-type BC with $-(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} \approx 0$ or as implicit OBC (cf. Sect. 8.5.3).

The heat transfer coefficients $\Phi_T^\dagger$ and Φ_T appearing in the Cauchy-type BC's (13.4) and (13.5), respectively, can be expressed by thermal resistances of composite materials in the form (exemplified for Φ_T according to (6.48), cf. Sect. 6.3.3.3)

$$
\Phi_T = \frac{1}{S \sum_i R_i}
\tag{13.9}
$$

where S is a specific exchange area and R_i is the specific thermal resistance of material i, for which typical cases are described in Appendix E.

The Cauchy-type BC can also be utilized to model *radiative heat transfer* on heated (solid) surfaces. The boundary heat flux for this type of thermal radiation can be given by Kaviany [305]

$$-(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} = F\sigma(T^4 - T_0^4) \tag{13.10}$$

where σ is the Stefan-Boltzmann constant and F is a form factor. The form factor F is related to the boundary emissivity and the position of the boundary relative to surrounding surfaces. The expression of thermal radiation (13.10) leads to a nonlinear heat transfer coefficient in the form

$$\Phi_T = F\sigma(T + T_0)(T^2 + T_0^2) \tag{13.11}$$

This type of radiation Cauchy-type BC is appropriate when a surface radiates to a black body environment that can be characterized by a single temperature.

The essential parameters required for solving (13.1) and (13.2) with (13.4)–(13.6) are listed in Table I.16 of Appendix I . *Steady-state* heat transport conditions occur if $\partial T/\partial t$ approaches to zero.[2]

13.2.2 Horizontal 2D Problems

The basic equations for the essentially horizontal, vertically averaged heat transport in unconfined and confined aquifers have been developed in Sect. 3.10.7 and summarized in Tables 3.10 and 3.11, respectively. The following 2D depth-integrated heat transport equations result

$$\frac{\partial}{\partial t}\Big[B\big(\varepsilon\rho c + (1-\varepsilon)\rho^s c^s\big)(T - T_0)\Big] + \nabla\cdot(\rho c\bar{\boldsymbol{q}}(T - T_0)) - \nabla\cdot(\bar{\boldsymbol{\Lambda}}\cdot\nabla T) = \bar{Q}_T + \bar{Q}_{Tw} \tag{13.12}$$

written in the divergence form and

$$B\big(\varepsilon\rho c + (1-\varepsilon)\rho^s c^s\big)\frac{\partial T}{\partial t} + \rho c\bar{\boldsymbol{q}}\cdot\nabla T - \nabla\cdot(\bar{\boldsymbol{\Lambda}}\cdot\nabla T) = \bar{Q}_T + \bar{Q}_{Tw} - \rho c(T - T_0)\bar{Q} \tag{13.13}$$

written in the convective form, which are associated with the constitutive relations

[2] Optionally, FEFLOW suppresses the time derivative term $\partial T/\partial t$ for solving steady-state solutions. A specific option exists, named *steady flow – transient transport*, in which the advective flow vector $\boldsymbol{q}$ is invariant with time.

$$
\begin{aligned}
B &= \begin{cases} h - f^B & \text{unconfined} \\ f^T - f^B & \text{confined} \end{cases} \\
\bar{\boldsymbol{\Lambda}} &= \bar{\boldsymbol{\Lambda}}_0 + \bar{\boldsymbol{\Lambda}}_0^s + \rho c \bar{\boldsymbol{D}}_{\text{mech}} \\
\bar{\boldsymbol{\Lambda}}_0 &= B\varepsilon\Lambda\boldsymbol{\delta} \qquad \text{isotropic thermal conductivity of liquid} \\
\bar{\boldsymbol{\Lambda}}_0^s &= \begin{cases} B(1-\varepsilon)\Lambda^s\boldsymbol{\delta} & \text{isotropic thermal conductivity of solid} \\ B(1-\varepsilon)\Lambda^s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \varXi_{\text{aniso}}^{\Lambda} \end{pmatrix} & \text{anisotropic thermal conductivity of solid in 3D} \end{cases} \\
\bar{\boldsymbol{D}}_{\text{mech}} &= \beta_T \|\bar{\boldsymbol{q}}\|\boldsymbol{\delta} + (\beta_L - \beta_T)\frac{\bar{\boldsymbol{q}} \otimes \bar{\boldsymbol{q}}}{\|\bar{\boldsymbol{q}}\|} \\
\bar{Q}_T &= B(\rho H^\star + \rho^s H_s^\star)
\end{aligned} \tag{13.14}
$$

where similarly the source/sink term $\bar{H}_e = B(\rho H + \rho^s H_s) = \bar{Q}_T + \bar{Q}_{Tw}$ is suitably split into the depth-integrated supply term $\bar{Q}_T = B(\rho H^\star + \rho^s H_s^\star)$ and depth-integrated well-type SPC term $\bar{Q}_{Tw}$. The solution of (13.12) or (13.13) for the temperature T is associated with the following BC's of Dirichlet, Neumann and Cauchy type as well as well-type SPC

$$
\begin{aligned}
T &= T_D & &\text{on} \quad \varGamma_{D_T} \times t[t_0, \infty) \\
\big((T - T_0)\rho c\bar{\boldsymbol{q}} - \bar{\boldsymbol{\Lambda}} \cdot \nabla T\big) \cdot \boldsymbol{n} &= \bar{q}_T^\dagger & &\text{on} \quad \varGamma_{N_T} \times t[t_0, \infty) \\
\big((T - T_0)\rho c\bar{\boldsymbol{q}} - \bar{\boldsymbol{\Lambda}} \cdot \nabla T\big) \cdot \boldsymbol{n} &= -\bar{\varPhi}_T^\dagger (T_C - T) & &\text{on} \quad \varGamma_{C_T} \times t[t_0, \infty) \\
\bar{Q}_{Tw} &= -\textstyle\sum_w (T_w - T_0)\rho c & &\text{on} \quad \boldsymbol{x}_w \in \varOmega \times t[t_0, \infty) \\
&\quad Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w)
\end{aligned} \tag{13.15}
$$

written for the divergence form of the heat transport equation and

$$
\begin{aligned}
T &= T_D & &\text{on} \quad \varGamma_{D_T} \times t[t_0, \infty) \\
-(\bar{\boldsymbol{\Lambda}} \cdot \nabla T) \cdot \boldsymbol{n} &= \bar{q}_T & &\text{on} \quad \varGamma_{N_T} \times t[t_0, \infty) \\
-(\bar{\boldsymbol{\Lambda}} \cdot \nabla T) \cdot \boldsymbol{n} &= -\bar{\varPhi}_T (T_C - T) & &\text{on} \quad \varGamma_{C_T} \times t[t_0, \infty) \\
\bar{Q}_{Tw} &= -\textstyle\sum_w (T_w - T)\rho c Q_w(t)\delta(\boldsymbol{x} - \boldsymbol{x}_w) & &\text{on} \quad \boldsymbol{x}_w \in \varOmega \times t[t_0, \infty)
\end{aligned} \tag{13.16}
$$

written for the convective form of the heat transport equation, imposed on $\varGamma = \varGamma_{C_T} \cup \varGamma_{N_T} \cup \varGamma_{C_T}$ and with the IC of the form

$$
T(\boldsymbol{x}, t_0) = T_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\varOmega} \tag{13.17}
$$

The essential parameters required for solving (13.12) and (13.13) with (13.15), (13.16) and (13.17) are listed in Table I.17 of Appendix I.

13.3 Finite Element Formulation

The governing ADE's for heat transport are mathematically similar to the paradigmatic ADE of a scalar quantity used in Chap. 8 to describe the fundamental concepts of FEM. Based on the principles given in Chap. 8 we use now the GFEM to solve the governing heat transport equations (13.1) and (13.2) subject to the corresponding BC's (13.4), (13.5) and IC (13.6). Since most of the details are equivalent to the ADE developments given in Chap. 8 we shall focus here only on the specific aspects of heat transport. For convenience we restrict our developments to 3D, vertical 2D and axisymmetric heat transport problems (Sect. 13.2.1). The formulations for the horizontal 2D heat transport in unconfined and confined aquifers (Sect. 13.2.2) will appear rather similar and can be easily deduced from the given statements.

13.3.1 Weak Forms

In analogy to the statements (8.48) and (8.55) of Sect. 8.5 we find the corresponding weak forms for the governing heat transport equation written in the divergence form (13.1) as

$$\begin{aligned}\int_{\Omega} w\frac{\partial}{\partial t}\Big[\big(\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s\big)(T-T_0)\Big]d\Omega - \int_{\Omega}(T-T_0)\rho c\boldsymbol{q}\cdot\nabla w d\Omega +\\ \int_{\Omega}\nabla w\cdot(\boldsymbol{\Lambda}\cdot\nabla T)d\Omega - \int_{\Omega} wQ_T d\Omega + \sum_{w} w(\boldsymbol{x}_w)(T_w-T_0)\rho c Q_w(t) +\\ \int_{\Gamma_{N_T}} wq_T^{\dagger}d\Gamma - \int_{\Gamma_{C_T}} w\Phi_T^{\dagger}(T_C-T)d\Gamma = 0, \quad \forall w \in H_0^1(\Omega)\end{aligned} \tag{13.18}$$

and written in the convective form (13.2) as

$$\begin{aligned}\int_{\Omega} w\big(\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s\big)\frac{\partial T}{\partial t}d\Omega + \int_{\Omega} w\rho c\boldsymbol{q}\cdot\nabla T d\Omega + \int_{\Omega}\nabla w\cdot(\boldsymbol{\Lambda}\cdot\nabla T)d\Omega +\\ \int_{\Omega} w[\rho c Q_h(T-T_0) - Q_T]d\Omega + \sum_{w} w(\boldsymbol{x}_w)(T_w-T)\rho c Q_w(t) +\\ \int_{\Gamma_{N_T}} wq_T d\Gamma - \int_{\Gamma_{C_T}} w\Phi_T(T_C-T)d\Gamma = 0, \quad \forall w \in H_0^1(\Omega)\end{aligned} \tag{13.19}$$

where w is a suitable weighting function and the boundary integrals are suitably separated into their segments $\Gamma = \Gamma_{D_T} \cup \Gamma_{N_T} \cup \Gamma_{C_T}$ imposed by the Dirichlet, Neumann and Cauchy-type BC's (13.4) and (13.5). OBC on $\mathit{\Gamma}_{N_O} \subset \mathit{\Gamma}_{N_T}$ represents special implementations of Neumann-type BC.[3]

13.3.2 GFEM and Resulting Matrix System

In using the FEM the unknown temperature variable T appearing in the weak statements (13.18) and (13.19) is replaced by a *continuous approximation* that assumes the separability of space and time (see Sect. 8.4). Thus

$$T(\boldsymbol{x}, t) \approx \sum_j N_j(\boldsymbol{x}) T_j(t), \quad j = 1, \ldots, N_\mathrm{P} \tag{13.20}$$

where j designates global nodal indices. Using the Galerkin method with the weighting function

$$w \to w_i = N_i, \quad i = 1, \ldots, N_\mathrm{P} \tag{13.21}$$

and applying the approximate solutions (13.20) in (13.18) and (13.19), the following matrix system of N_P equations (cf. Sect. 8.9) results

$$\boldsymbol{P} \cdot \dot{\boldsymbol{T}} + \boldsymbol{L} \cdot \boldsymbol{T} - \boldsymbol{W} = \boldsymbol{0} \tag{13.22}$$

[3] A boundary with OBC on $\mathit{\Gamma}_{N_O}$ can be separated from the Neumann boundary $\mathit{\Gamma}_{N_T}$ so that for the divergence form

$$\int_{\mathit{\Gamma}_{N_T}} w q_T^\dagger d\mathit{\Gamma} = \int_{\mathit{\Gamma}_{N_T} \setminus \mathit{\Gamma}_{N_O}} w q_T^\dagger d\mathit{\Gamma} + \int_{\mathit{\Gamma}_{N_O}} w\big((T - T_0)\rho c \boldsymbol{q} - \boldsymbol{\Lambda} \cdot \nabla T\big) \cdot \boldsymbol{n} d\mathit{\Gamma}$$

and for the convective form

$$\int_{\mathit{\Gamma}_{N_T}} w q_T d\mathit{\Gamma} = \int_{\mathit{\Gamma}_{N_T} \setminus \mathit{\Gamma}_{N_O}} w q_T d\mathit{\Gamma} - \int_{\mathit{\Gamma}_{N_O}} w(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} d\mathit{\Gamma}$$

The implicit treatment of OBC requires the incorporation of the $\mathit{\Gamma}_{N_O}$—integrals into the LHS of the resulting matrix system (see below). In contrast, a natural Neumann-type BC with $-(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} \approx 0$ on $\mathit{\Gamma}_{N_O}$ is often the preferred alternative formulation for an OBC. Note, however, that for both cases in the divergence form the boundary flux $\boldsymbol{q} \cdot \boldsymbol{n}$ must be known a priori. The boundary flux $\boldsymbol{q} \cdot \boldsymbol{n}$ can be either explicitly given from a Neumann-type BC $q_h = \boldsymbol{q} \cdot \boldsymbol{n}$ for flow or must be computed by a postprocessing budget evaluation of the flow equation on the corresponding outflowing boundary section imposed by Dirichlet-type or Cauchy-type BC of flow.

where

$$\boldsymbol{T} = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_{N_\mathrm{P}} \end{pmatrix}, \quad \dot{\boldsymbol{T}} = \begin{pmatrix} \frac{dT_1}{dt} \\ \frac{dT_2}{dt} \\ \vdots \\ \frac{dT_{N_\mathrm{P}}}{dt} \end{pmatrix} \tag{13.23}$$

and the matrices and RHS vector

$$\begin{aligned}
\boldsymbol{P} = H_{ij} &= \sum_e \int_{\Omega^e} \big(\varepsilon^e s^e \rho^e c^e + (1-\varepsilon^e)\rho^{s\,e} c^{s\,e}\big) N_i N_j \, d\Omega^e \\
\boldsymbol{L} = L_{ij} &= \begin{cases} \sum_e \Big(-\int_{\Omega^e} \rho^e c^e \boldsymbol{q}^e \cdot \nabla N_i N_j \, d\Omega^e + \int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{\Lambda}^e \cdot \nabla N_j) d\Omega^e + \\ \quad \int_{\Gamma^e_{C_T}} \Phi_T^{\dagger e} N_i N_j \, d\Gamma^e + \int_{\Gamma^e_{N_O}} N_i (\rho^e c^e \boldsymbol{q}^e N_j - \boldsymbol{\Lambda}^e \cdot \nabla N_j) \cdot \boldsymbol{n} \, d\Gamma^e \Big) \\ \hfill \text{divergence form} \\ \sum_e \Big(\int_{\Omega^e} N_i \rho^e c^e \boldsymbol{q}^e \cdot \nabla N_j \, d\Omega^e + \int_{\Omega^e} \nabla N_i \cdot (\boldsymbol{\Lambda}^e \cdot \nabla N_j) d\Omega^e + \\ \quad \int_{\Omega^e} \rho^e c^e Q_h^e N_i N_j \, d\Omega^e + \int_{\Gamma^e_{C_T}} \Phi_T^e N_i N_j \, d\Gamma^e - \\ \quad \int_{\Gamma^e_{N_O}} N_i (\boldsymbol{\Lambda}^e \cdot \nabla N_j) \cdot \boldsymbol{n} \, d\Gamma^e \Big) - \delta_{ij} \rho c Q_w(t)\big|_i \\ \hfill \text{convective form} \end{cases} \\
\boldsymbol{W} = W_i &= \begin{cases} \sum_e \Big(-\int_{\Omega^e} \rho^e c^e \boldsymbol{q}^e \cdot \nabla N_i T_0 \, d\Omega^e + \int_{\Omega^e} N_i Q_T^e \, d\Omega^e + \\ \quad \int_{\Gamma^e_{C_T}} N_i \Phi_T^{\dagger e} T_C^e \, d\Gamma^e + \int_{\Gamma^e_{N_O}} N_i \rho^e c^e T_0 \, \boldsymbol{q}^e \cdot \boldsymbol{n} \, d\Gamma^e - \\ \quad \int_{\Gamma^e_{N_T} \setminus \Gamma^e_{N_O}} N_i q_T^{\dagger e} \, d\Gamma^e \Big) - (T_w - T_0) \rho c Q_w(t)\big|_i \quad \text{divergence form} \\ \sum_e \Big(\int_{\Omega^e} N_i (Q_T^e + \rho^e c^e Q_h^e T_0) d\Omega^e + \int_{\Gamma^e_{C_T}} N_i \Phi_T^e T_C^e \, d\Gamma^e - \\ \quad \int_{\Gamma^e_{N_T} \setminus \Gamma^e_{N_O}} N_i q_T^e \, d\Gamma^e \Big) - T_w \rho c Q_w(t)\big|_i \quad \text{convective form} \end{cases}
\end{aligned} \tag{13.24}$$

where $(i, j = 1, \ldots, N_\mathrm{P})$ and $(e = 1, \ldots, N_\mathrm{E})$. Note that we assumed in the divergence form that the spatial and temporal derivative terms related to the reference temperature T_0 are negligible. The integrals appearing in (13.24) are integrated on element level in the local coordinates as described in Sect. 8.12. Analytical evaluations of partial integral terms of (13.24) can be deduced from developments done in Appendix H for selected element types. The differential elements $d\Omega^e$ and $d\Gamma^e$ differ for 3D, 2D and axisymmetric problems as given by (8.122)–(8.124), respectively. Is is important to note that the resulting global system of equations (13.22) is *unsymmetric* since the matrix $\boldsymbol{L}$ is unsymmetric due to advection.

For advective-dominant heat transport the discretized system (13.22) can be easily combined with upwind strategies as introduced in Sect. 8.14. Useful upwind strategies refer to the SU and FU methods (Sect. 8.14.3), SC method (Sect. 8.14.4) and PGLS method (Sect. 8.14.5), in which the tensor of mechanical dispersion $\boldsymbol{D}_{\text{mech}}$ as part of the hydrodynamic thermodispersion tensor $\boldsymbol{\Lambda}$ is appropriately modified by stabilization terms in dependence on the actual spatial and temporal discretizations or temperature gradients. The required modifications of $\boldsymbol{D}^{e}_{\text{mech}}$ for each element e were discussed in the preceding Sect. 11.6.3 and summarized in Table 11.3.

13.3.3 Time Integration

The resulting matrix system (13.22) has to be solved in time t with the associated IC's via suitable single-step semi-implicit or fully implicit time marching recurrence schemes as described in Sect. 8.13. The GLS predictor-corrector time stepping method combined with an automatic error-controlled time step selection strategy is usually preferred. Its solution steps applied to the heat transport are fully equivalent to the procedures as thoroughly described above in Sect. 8.13.5 (summarized in Table 8.7) for a general ADE, in Sect. 10.7.5 for unsaturated flow, in Sect. 11.6.4 for density-variable flow, mass and heat transport and in Sect. 12.3.3 for reactive mass transport. In the context of heat transport the (corrector) recurrence scheme reads

$$\Big(\frac{\boldsymbol{P}}{\theta\Delta t_n}+\boldsymbol{L}\Big)\cdot\boldsymbol{T}_{n+1}=\boldsymbol{P}\cdot\Big[\frac{\boldsymbol{T}_n}{\theta\Delta t_n}+\big(\tfrac{1}{\theta}-1\big)\dot{\boldsymbol{T}}_n\Big]+\boldsymbol{W}_{n+1} \tag{13.25}$$

to solve $\boldsymbol{T}_{n+1}$ at the new time plane $n+1$, where $\theta\in(\frac{1}{2},1)$ for the TR and BE scheme, respectively. On the other hand, for user-defined (fixed) time step sizes Δt_n the θ−method (Sect. 8.13.4) is applicable

$$\Big(\frac{\boldsymbol{P}}{\Delta t_n}+\boldsymbol{L}\theta\Big)\cdot\boldsymbol{T}_{n+1}=\Big(\frac{\boldsymbol{P}}{\Delta t_n}-\boldsymbol{L}(1-\theta)\Big)\cdot\boldsymbol{T}_n+\big(\boldsymbol{W}_{n+1}\theta+\boldsymbol{W}_n(1-\theta)\big) \tag{13.26}$$

where $\theta\in(\frac{1}{2},\frac{2}{3},1)$ for the Crank-Nicolson, the Galerkin-in-time and the fully implicit scheme, respectively.

13.4 Heat Budget Analysis

The CBFM is used, as introduced in Sect. 8.19.2, to obtain a precise heat budget analysis. It is based on the specific weak formulations of the governing heat transport equations. The corresponding boundary heat fluxes on Γ have to be evaluated from the basic weak statements (13.18) and (13.19) of the divergence and convective form, respectively, written as

$$\begin{aligned}\int_\Gamma N_i\, q^\dagger_{n_T}\, d\Gamma = &-\int_\Omega N_i \frac{\partial}{\partial t}\Big[\big(\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s\big)(T-T_0)\Big] d\Omega + \\ &\int_\Omega (T-T_0)\rho c\boldsymbol{q}\cdot\nabla N_i d\Omega - \int_\Omega \nabla N_i\cdot(\boldsymbol{\Lambda}\cdot\nabla T)d\Omega + \int_\Omega N_i Q_T d\Omega - \\ &(T_w - T_0)\rho c Q_w(t)|_i\end{aligned} \tag{13.27}$$

$$\begin{aligned}\int_\Gamma N_i\, q_{n_T}\, d\Gamma = &-\int_\Omega N_i\big(\varepsilon s\rho c + (1-\varepsilon)\rho^s c^s\big)\frac{\partial T}{\partial t} d\Omega - \int_\Omega N_i\rho c\boldsymbol{q}\cdot\nabla T d\Omega - \\ &\int_\Omega \nabla N_i\cdot(\boldsymbol{\Lambda}\cdot\nabla T)d\Omega - \int_\Omega N_i[\rho c Q_h(T-T_0) - Q_T]d\Omega - \\ &(T_w - T)\rho c Q_w(t)|_i\end{aligned} \tag{13.28}$$

to compute $q^\dagger_{n_T}$ or q_{n_T}, where T is known at evaluation time t_{n+1}. Note that the boundary heat flux $q^\dagger_{n_T} = \big((T-T_0)\rho c\boldsymbol{q} - \boldsymbol{D}_k\cdot\nabla T\big)\cdot\boldsymbol{n}$ of the divergence form encompasses the total heat flux consisting of the advective and dispersive parts, while the boundary heat flux $q_{n_T} = -(\boldsymbol{\Lambda}\cdot\nabla T)\cdot\boldsymbol{n}$ of the convective form consists only of the dispersive part. Accordingly, for the convective form an additional balance expression of the missing advective part $q^a_{n_T} = (T-T_0)\rho c\boldsymbol{q}\cdot\boldsymbol{n}$ to obtain $q^\dagger_{n_T} = q_{n_T} + q^a_{n_T}$ is needed. This is attained by using an auxiliary weak formulation applied to the governing flow equation (10.5) as described in Sect. 8.19.2.4. We find

$$\begin{aligned}\int_\Gamma N_i\, q^a_{n_T}\, d\Gamma = &-\int_\Omega \rho c\nabla N_i\cdot[k_r\boldsymbol{K}f_\mu\cdot(\nabla h + \chi\boldsymbol{e})](T-T_0)d\Omega - \\ &\int_\Omega \rho c N_i\nabla T\cdot[k_r\boldsymbol{K}f_\mu\cdot(\nabla h + \chi\boldsymbol{e})]d\Omega + \\ &\int_\Omega \rho c N_i(T-T_0)(Q_h + Q_{hw} + Q_{\mathrm{EOB}})d\Omega - \\ &\int_\Omega \rho c N_i(T-T_0)\Big(sS_o\frac{\partial h}{\partial t} + \varepsilon\frac{\partial s}{\partial t}\Big)d\Omega\end{aligned} \tag{13.29}$$

to compute $q^a_{n_T}$, where h, s and T are known at evaluation time t_{n+1}. Expanding the boundary flux on Γ as described in Sect. 8.19.2 the following matrix system results to solve the consistent boundary total heat flux vector $\boldsymbol{q}^\dagger_{n_T}$, viz.,

$$\begin{aligned}\boldsymbol{M}\cdot\boldsymbol{q}^\dagger_{n_T} = &-\boldsymbol{P}\cdot\dot{\boldsymbol{T}} - \boldsymbol{L}^\dagger\cdot\boldsymbol{T} + \boldsymbol{W}^\dagger \\ &-\begin{cases}\boldsymbol{0} & \text{divergence form}\\ \boldsymbol{V}(\boldsymbol{h})\cdot(\boldsymbol{T}-\boldsymbol{T}_0) + \boldsymbol{A}(\boldsymbol{T})\cdot\boldsymbol{h} - \boldsymbol{F}(\boldsymbol{T},\boldsymbol{s},\dot{\boldsymbol{h}},\dot{\boldsymbol{s}}) & \text{convective form}\end{cases}\end{aligned} \tag{13.30}$$

with known $\boldsymbol{T}$, $\boldsymbol{T}_0$, $\dot{\boldsymbol{T}}$, $\boldsymbol{h}$, $\dot{\boldsymbol{h}}$, $\boldsymbol{s}$ and $\dot{\boldsymbol{s}}$ at the corresponding evaluation time t_{n+1}, where $\boldsymbol{P}$ is defined in (13.24) and

$$\begin{aligned}
\boldsymbol{M} = M_{ij} &= \int_{\Gamma} N_i N_j d\Gamma \\
\boldsymbol{L}^{\dagger} = L_{ij}^{\dagger} &= \begin{cases} -\int_{\Omega} \rho c \boldsymbol{q} \cdot \nabla N_i N_j d\Omega + \int_{\Omega} \nabla N_i \cdot (\boldsymbol{\Lambda} \cdot \nabla N_j) d\Omega & \\ & \text{divergence form} \\ \int_{\Omega} N_i \rho c \boldsymbol{q} \cdot \nabla N_j d\Omega + \int_{\Omega} \nabla N_i \cdot (\boldsymbol{\Lambda} \cdot \nabla N_j) d\Omega + & \\ \int_{\Omega} \rho c Q_h N_i N_j d\Omega - \delta_{ij} \rho c Q_w(t)\big|_i & \text{convective form} \end{cases} \\
\boldsymbol{W}^{\dagger} = W_i^{\dagger} &= \begin{cases} -\int_{\Omega} \rho c \boldsymbol{q} \cdot \nabla N_i T_0 d\Omega + \int_{\Omega} N_i Q_T d\Omega - (T_w - T_0) \rho c Q_w(t)\big|_i & \\ & \text{divergence form} \\ \int_{\Omega} N_i (Q_T + \rho c Q_h T_0) d\Omega - T_w \rho c Q_w(t)\big|_i & \text{convective form} \end{cases} \\
\boldsymbol{V} = V_{ij} &= \int_{\Omega} \rho c N_i \nabla N_j \cdot [k_r \boldsymbol{K} f_{\mu} \cdot (\nabla h + \chi \boldsymbol{e})] d\Omega \\
\boldsymbol{A} = A_{ij} &= \int_{\Omega} \rho c \nabla N_i \cdot [k_r \boldsymbol{K} f_{\mu} \cdot (\nabla N_j + \chi \boldsymbol{e})](T - T_0) d\Omega \\
\boldsymbol{F} = F_i &= \int_{\Omega} \rho c N_i (T - T_0) \big(Q_h + Q_{\text{EOB}} - s S_o \tfrac{\partial h}{\partial t} - \varepsilon \tfrac{\partial s}{\partial t} \big) d\Omega - \\
&\quad (T - T_0) \rho c Q_w(t)\big|_i
\end{aligned} \tag{13.31}$$

in which $(i, j = 1, \ldots, N_{\text{P}})$ and $(e = 1, \ldots, N_{\text{E}})$. Note that the spatial and temporal derivative terms in the divergence form related to the reference temperature T_0 are again neglected. Furthermore, note that $\boldsymbol{V}$, $\boldsymbol{A}$ and $\boldsymbol{F}$ are only needed for the convective form. In the budget analysis the integral boundary balance flux $\boldsymbol{Q}_{n_T}$ is directly evaluated at each boundary node by

$$\begin{aligned}
\boldsymbol{Q}_{n_T} &= -\boldsymbol{M} \cdot \boldsymbol{q}_{n_T}^{\dagger} \\
&= \boldsymbol{P} \cdot \dot{\boldsymbol{T}} + \boldsymbol{L}^{\dagger} \cdot \boldsymbol{T} - \boldsymbol{W}^{\dagger} \\
&\quad + \begin{cases} \boldsymbol{0} & \text{divergence form} \\ \boldsymbol{V}(\boldsymbol{h}) \cdot (\boldsymbol{T} - \boldsymbol{T}_0) + \boldsymbol{A}(\boldsymbol{T}) \cdot \boldsymbol{h} - \boldsymbol{F}(\boldsymbol{T}, \boldsymbol{s}, \dot{\boldsymbol{h}}, \dot{\boldsymbol{s}}) & \text{convective form} \end{cases}
\end{aligned} \tag{13.32}$$

where $\boldsymbol{Q}_{n_T}$ corresponds to the nodal vector of the integral boundary heat flux.

13.5 Incorporation of Borehole Heat Exchangers (BHE's)

13.5.1 Introduction

In shallow aquifers a modern geothermal heat extraction technology (*geoexchange*) concerns the use of borehole heat exchanger (BHE) systems of different construction. The most common in practice are single U-shape pipe (consisting of an inlet pipe, an outlet pipe and grout), double U-shape pipe (consisting of two inlet pipes, two outlet pipes and grout) and coaxial pipe (consisting of an inlet pipe included

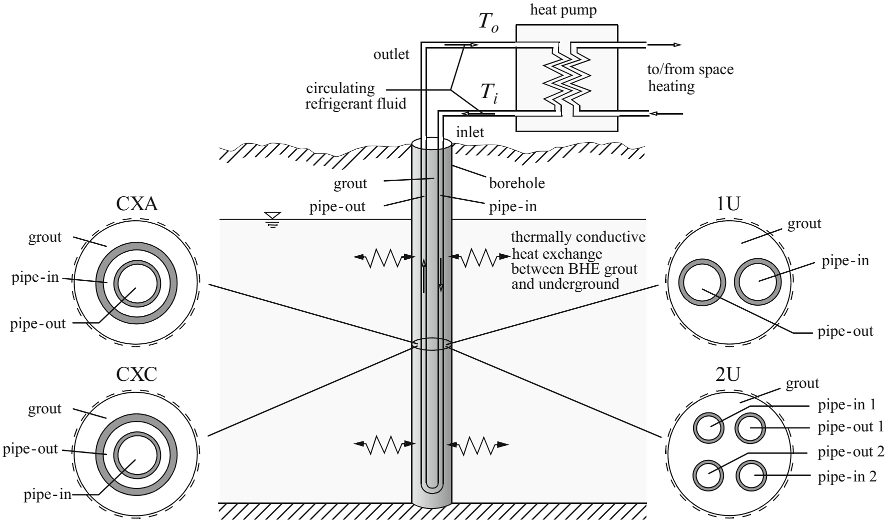


Fig. 13.1 Closed-loop scheme of BHE with different configurations in form of single U-shape pipe (1U), double U-shape pipe (2U), coaxial pipe with annular inlet (CXA) and coaxial pipe with centered inlet (CXC)

with an outlet pipe and grout) installations. Such heat exchangers form a vertical borehole system, where a refrigerant of a heat pump circulates in closed pipes (*closed loop system*). These pipes inserted vertically in a borehole are fixed by filling the borehole with some sort of grout (backfill) material. It is in contact with the surrounding soil, where conductive-convective heat transfer processes occur (Fig. 13.1).

The modeling and simulation of the complex transient 3D transport phenomena of BHE's is complicated and cumbersome due to the extreme geometric aspect ratios (extreme slenderness). A number of design tools based on finite element or finite volume codes were used in the past to develop fully discretized BHE models which are able to account for transient effects as well as for the correct borehole geometry [28,330,477]. To reduce the computational effort some of the models were limited to a 2D description [14,510,577]. However, if a complete description of the borehole geometry is needed, only 3D models can include vertical heat transport inside and outside the borehole, different ground layers, the vertical gradient of the undisturbed ground temperature, the transient fluid transport inside the pipes, the thermal short-circuiting between the upward and downward pipes and the correct BC's at the upper and lower boundary. On the other hand, the main disadvantage of fully discretized 3D models is that, even on modern and powerful computers and despite the possibility of parallel computing, extensive computation times result due to the high number of small elements needed for a suited discretization of the borehole cross-sections.

Accordingly, the extreme geometric aspect ratios require more advanced and efficient numerical strategies [145, 146], in particular

- The analytical BHE method based on Eskilson and Claesson's solution [159] and
- The numerical BHE method based on Al-Khoury et al.'s solution [6–8].

Often, the local processes within BHE can suitably be modeled via the analytical technique under the major assumption that local steady-state conditions are considered, where a thermal equilibrium immediately occurs between inlet and outlet pipes for a given solid temperature at the borehole wall. Such type of analytical solutions has been firstly introduced by Eskilson and Claesson [159]. Their local analytical model can be taken as a robust and efficient alternative to the more general Al-Khoury et al.'s numerical strategy, particularly for long-term predictions. Eskilson and Claesson's analytical solution has been extended to different types of BHE and embedded in a general iterative finite-element strategy for solving the overall problem in Diersch et al. [145, 146]. While the Al-Khoury et al.'s numerical approach has proven appropriated over the full time range of processes, Eskilson and Claesson's analytical solution is not suited for short-term predictions (say, thermal responses in a time range smaller than some hours), however, for long-term predictions the analytical solution has been shown in a well and reasonable accuracy in comparison to the general Al-Khoury et al.'s numerical solution [146].

For both BHE solution techniques an improved pipe-to-grout approximation method will be preferred which is based on the extension given by the so-called *thermal resistance and capacity model* (TRCM) introduced by Bauer et al. [29]. Previous formulations of thermal resistances, such as provided in [6–8], use only one single capacity point for the grout material. However, Bauer et al. [29] have shown that such a single grout point approximation is insufficient and less accurate for transient computations. The TRCM takes the capacity of the grouting material with one capacity per pipe into account and has proven accurate and effective both for transient and steady-state BHE conditions, see [29] for more.

13.5.2 Implementation of BHE's

The aquifer is discretized in FEFLOW by using 3D prismatic finite elements, where BHE systems are modeled by vertical boreholes. Each borehole is discretized by a number of N_{BHE} nodes, which are linked to 1D pipe elements as exemplified in Fig. 13.2 for a single 2U exchanger borehole. The N_{BHE} nodes represent inner boundary nodes of the soil (porous medium) s, to which the heat exchange is performed between the soil and the BHE. The detailed pipe-grout structure of each BHE remains completely hidden in the global mesh of the aquifer, where each BHE is viewed as a singular well-type (SPC) representation of N_{BHE} nodes.

The heat exchange mechanisms between the soil and the BHE's lead to additional heat transfer terms appearing in the discretized heat transport equations for the porous medium (13.22) on the RHS as

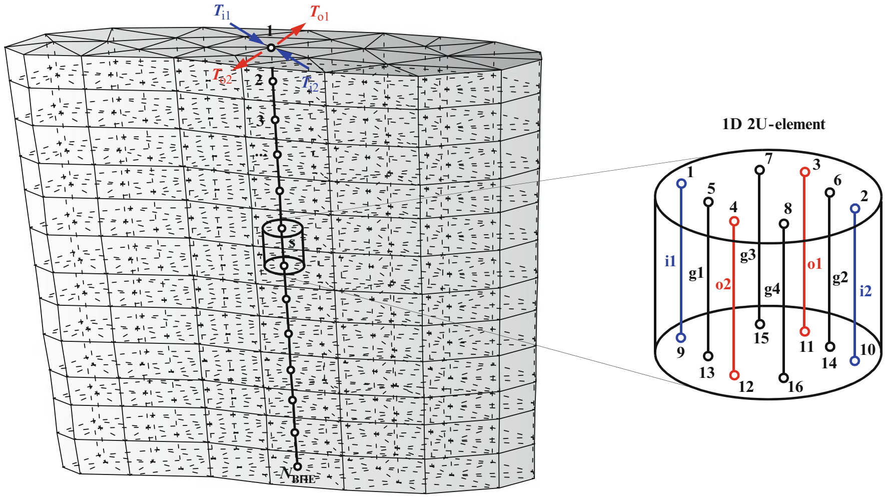


Fig. 13.2 Discretized single 2U exchanger borehole

$$\boldsymbol{P}^s \cdot \dot{\boldsymbol{T}}^s + \boldsymbol{L}^s \cdot \boldsymbol{T}^s = \boldsymbol{W}^s \underbrace{-\hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^\pi) + \boldsymbol{R}^\pi \cdot \boldsymbol{T}^s}_{\text{additional BHE exchange terms}} \tag{13.33}$$

and introduce an extra matrix system containing the heat transport equations of BHE's written in the form

$$\boldsymbol{P}^\pi \cdot \dot{\boldsymbol{T}}^\pi + \boldsymbol{L}^\pi \cdot \boldsymbol{T}^\pi = \boldsymbol{W}^\pi - \hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^s) \tag{13.34}$$

where superscript s indicates the soil equations, superscript π designates the internal BHE (pipe-grout) equations for heat, the vectors $\hat{\boldsymbol{R}}^{s\pi}$, $\hat{\boldsymbol{R}}^{\pi s}$ and matrix $\boldsymbol{R}^\pi$ represent BHE-soil heat transfer relations as derived in Appendix M. The soil and BHE equations (13.33) and (13.34), respectively, are coupled via the heat transfer terms and can be written in the compact form as

$$\begin{pmatrix} \boldsymbol{P}^\pi & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}^s \end{pmatrix} \cdot \begin{pmatrix} \dot{\boldsymbol{T}}^\pi \\ \dot{\boldsymbol{T}}^s \end{pmatrix} + \begin{pmatrix} \boldsymbol{L}^\pi & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{L}^s - \boldsymbol{R}^\pi \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{T}^\pi \\ \boldsymbol{T}^s \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}^\pi - \hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^s) \\ \boldsymbol{W}^s - \hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^\pi) \end{pmatrix} \tag{13.35}$$

for solving the internal (local) BHE temperatures $\boldsymbol{T}^\pi = \boldsymbol{T}^\pi(\boldsymbol{x}, t)$ and the global soil temperatures $\boldsymbol{T}^s = \boldsymbol{T}^s(\boldsymbol{x}, t)$ in dependence on the chosen BHE method as follows.

13.5.2.1 Analytical BHE Solution

For the analytical BHE method the matrices and vectors in (13.35) simplify:

$$\boldsymbol{P}^\pi = \boldsymbol{0}, \quad \boldsymbol{W}^\pi = \boldsymbol{0}, \quad \boldsymbol{L}^\pi = \boldsymbol{\delta} \tag{13.36}$$

so that (13.35) decouples into

$$\begin{aligned} \boldsymbol{T}^{\pi} &= -\hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^{s}) \\ \boldsymbol{P}^{s}\cdot\dot{\boldsymbol{T}}^{s} + (\boldsymbol{L}^{s} - \boldsymbol{R}^{\pi})\cdot\boldsymbol{T}^{s} &= \boldsymbol{W}^{s} - \hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^{\pi}) \end{aligned} \tag{13.37}$$

where $-\hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^{s})$, $-\hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^{\pi})$ and $-\boldsymbol{R}^{\pi}$ are given by (M.116)–(M.118), respectively, in Appendix M and leads in the temporally discretized form (cf. Sect. 13.3.3) to

$$\begin{aligned} \boldsymbol{T}^{\pi}_{n+1} &= -\hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^{s}_{n+1}) \\ \boldsymbol{A}^{s}\cdot\boldsymbol{T}^{s}_{n+1} &= \boldsymbol{B}^{s}(\boldsymbol{T}^{s}_{n+1},\boldsymbol{T}^{s}_{n}) \end{aligned} \tag{13.38}$$

where

$$\begin{aligned} \boldsymbol{A}^{s} &= \tfrac{\boldsymbol{P}^{s}}{\theta\Delta t_n} + \boldsymbol{L}^{s} - \boldsymbol{R}^{\pi} \\ \boldsymbol{B}^{s}(\boldsymbol{T}^{s}_{n+1},\boldsymbol{T}^{s}_{n}) &= \boldsymbol{P}^{s}\cdot\left[\tfrac{\boldsymbol{T}^{s}_{n}}{\theta\Delta t_n} + \left(\tfrac{1}{\theta}-1\right)\dot{\boldsymbol{T}}^{s}_{n}\right] + \boldsymbol{W}^{s}_{n+1} - \hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^{\pi}_{n+1}(\boldsymbol{T}^{s}_{n+1})) \end{aligned} \tag{13.39}$$

for the corrector recurrence scheme and

$$\begin{aligned} \boldsymbol{A}^{s} &= \tfrac{\boldsymbol{P}^{s}}{\Delta t_n} + \theta(\boldsymbol{L}^{s} - \boldsymbol{R}^{\pi}) \\ \boldsymbol{B}^{s}(\boldsymbol{T}^{s}_{n+1},\boldsymbol{T}^{s}_{n}) &= \left(\tfrac{\boldsymbol{P}^{s}}{\Delta t_n} - (\boldsymbol{L}^{s} - \boldsymbol{R}^{\pi})(1-\theta)\right)\cdot\boldsymbol{T}^{s}_{n} + \\ &\quad \left(\boldsymbol{W}^{s}_{n+1} - \hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^{\pi}_{n+1}(\boldsymbol{T}^{s}_{n+1}))\right)\theta + \\ &\quad \left(\boldsymbol{W}^{s}_{n} - \hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^{\pi}_{n}(\boldsymbol{T}^{s}_{n}))\right)(1-\theta) \end{aligned} \tag{13.40}$$

for the $\theta-$method of time stepping.

Since in (13.38) the BHE temperatures $\boldsymbol{T}^{\pi}_{n+1} = -\hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^{s}_{n+1})$ are dependent on the soil temperature $\boldsymbol{T}^{s}_{n+1}$ by complex analytical expressions, the resulting matrix system $\boldsymbol{A}^{s}\cdot\boldsymbol{T}^{s}_{n+1} = \boldsymbol{B}^{s}(\boldsymbol{T}^{s}_{n+1},\boldsymbol{T}^{s}_{n})$ in (13.38) for solving the soil temperatures $\boldsymbol{T}^{s}_{n+1}$ becomes nonlinear and must be solved via the following iterative procedure

$$\begin{aligned} \text{starting solution } \tau = 0\text{:} &\quad \boldsymbol{A}^{s}\cdot\boldsymbol{T}^{s,\tau}_{n+1} = \boldsymbol{B}^{s}(\boldsymbol{T}^{s}_{n}) \\ \text{iteration } \tau+1\text{:} &\quad \boldsymbol{A}^{s}\cdot\boldsymbol{T}^{s,\tau+1}_{n+1} = \boldsymbol{B}^{s}(\boldsymbol{T}^{s,\tau}_{n+1},\boldsymbol{T}^{s}_{n}) \end{aligned} \tag{13.41}$$

where $\tau = 0, 1, 2, \ldots$ corresponds to an iteration counter. The iterations at the current time plane $n+1$ are stopped if

$$\|\boldsymbol{T}^{s,\tau+1}_{n+1} - \boldsymbol{T}^{s,\tau}_{n+1}\|_{L_p} \leq \epsilon \tag{13.42}$$

where ϵ is a given dimensionless error tolerance and $\|.\|_{L_p}$ designates a suitable (maximum or RMS) error norm.

13.5.2.2 Numerical BHE Solution

For the numerical BHE method the heat transfer matrices in (13.35) result (see Appendix M):

$$\hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^s) = \boldsymbol{R}^{\pi s} \cdot \boldsymbol{T}^s, \quad \hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^\pi) = \boldsymbol{R}^{s\pi} \cdot \boldsymbol{T}^\pi \tag{13.43}$$

so that (13.35) becomes

$$\begin{pmatrix} \boldsymbol{P}^\pi & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}^s \end{pmatrix} \cdot \begin{pmatrix} \dot{\boldsymbol{T}}^\pi \\ \dot{\boldsymbol{T}}^s \end{pmatrix} + \begin{pmatrix} \boldsymbol{L}^\pi & \boldsymbol{R}^{\pi s} \\ \boldsymbol{R}^{s\pi} & \boldsymbol{L}^s - \boldsymbol{R}^\pi \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{T}^\pi \\ \boldsymbol{T}^s \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}^\pi \\ \boldsymbol{W}^s \end{pmatrix} \tag{13.44}$$

where $\boldsymbol{P}^\pi$, $\boldsymbol{L}^\pi$, $(\boldsymbol{R}^{\pi s}, \boldsymbol{W}^\pi, \boldsymbol{T}^\pi)$, $\boldsymbol{R}^{s\pi}$ and $\boldsymbol{R}^\pi$ are given by (M.126)–(M.128), (M.132) and (M.133), respectively, in Appendix M and leads in the temporally discretized form (cf. Sect. 13.3.3) to

$$\begin{pmatrix} \boldsymbol{A}^\pi & \boldsymbol{R}^{\pi s} \\ \boldsymbol{R}^{s\pi} & \boldsymbol{A}^s \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{T}^\pi \\ \boldsymbol{T}^s \end{pmatrix}_{n+1} = \begin{pmatrix} \boldsymbol{B}^\pi \\ \boldsymbol{B}^s \end{pmatrix}_{n+1,n} \tag{13.45}$$

where

$$\begin{aligned} \boldsymbol{A}^\pi &= \tfrac{\boldsymbol{P}^\pi}{\theta \Delta t_n} + \boldsymbol{L}^\pi \\ \boldsymbol{B}^\pi &= \boldsymbol{P}^\pi \cdot \left[\tfrac{\boldsymbol{T}_n^\pi}{\theta \Delta t_n} + \left(\tfrac{1}{\theta} - 1\right) \dot{\boldsymbol{T}}_n^\pi \right] + \boldsymbol{W}_{n+1}^\pi \\ \boldsymbol{A}^s &= \tfrac{\boldsymbol{P}^s}{\theta \Delta t_n} + \boldsymbol{L}^s - \boldsymbol{R}^\pi \\ \boldsymbol{B}^s &= \boldsymbol{P}^s \cdot \left[\tfrac{\boldsymbol{T}_n^s}{\theta \Delta t_n} + \left(\tfrac{1}{\theta} - 1\right) \dot{\boldsymbol{T}}_n^s \right] + \boldsymbol{W}_{n+1}^s \end{aligned} \tag{13.46}$$

for the corrector recurrence scheme and

$$\begin{aligned} \boldsymbol{A}^\pi &= \tfrac{\boldsymbol{P}^\pi}{\Delta t_n} + \theta \boldsymbol{L}^\pi \\ \boldsymbol{B}^\pi &= \left(\tfrac{\boldsymbol{P}^\pi}{\Delta t_n} - \boldsymbol{L}^\pi (1-\theta) \right) \cdot \boldsymbol{T}_n^\pi + \boldsymbol{W}_{n+1}^\pi \theta + \boldsymbol{W}_n^\pi (1-\theta) \\ \boldsymbol{A}^s &= \tfrac{\boldsymbol{P}^s}{\Delta t_n} + \theta (\boldsymbol{L}^s - \boldsymbol{R}^\pi) \\ \boldsymbol{B}^s &= \left(\tfrac{\boldsymbol{P}^s}{\Delta t_n} - (\boldsymbol{L}^s - \boldsymbol{R}^\pi)(1-\theta) \right) \cdot \boldsymbol{T}_n^s + \boldsymbol{W}_{n+1}^s \theta + \boldsymbol{W}_n^s (1-\theta) \end{aligned} \tag{13.47}$$

for the $\theta-$method of time stepping.

Due to the heat transfer relations (13.43) the resulting matrix system (13.45) becomes basically linear. For the solution of (13.45) a *static condensation strategy* (also known as substructuring technique [590] frequently used in finite-element structural engineering) is preferred, where the internal BHE variables can be eliminated from (13.45). In doing so, the first row of the matrix system (13.45) reads

$$\boldsymbol{A}^\pi \cdot \boldsymbol{T}_{n+1}^\pi + \boldsymbol{R}^{\pi s} \cdot \boldsymbol{T}_{n+1}^s = \boldsymbol{B}_{n+1,n}^\pi \tag{13.48}$$

and yields

$$\boldsymbol{T}_{n+1}^{\pi} = (\boldsymbol{A}^{\pi})^{-1} \cdot \left(\boldsymbol{B}_{n+1,n}^{\pi} - \boldsymbol{R}^{\pi s} \cdot \boldsymbol{T}_{n+1}^{s}\right) \tag{13.49}$$

Taking the second row of (13.45) the BHE temperature vector $\boldsymbol{T}_{n+1}^{\pi}$ can be eliminated by using (13.49). It finally gives a reduced equation system of the following form

$$(\boldsymbol{A}^{s} - \boldsymbol{A}^{\pi s}) \cdot \boldsymbol{T}_{n+1}^{s} = \boldsymbol{B}_{n+1,n}^{s} - \boldsymbol{B}_{n+1,n}^{\pi s} \tag{13.50}$$

with

$$\begin{aligned} \boldsymbol{A}^{\pi s} &= \boldsymbol{R}^{s\pi} \cdot \left((\boldsymbol{A}^{\pi})^{-1} \cdot \boldsymbol{R}^{\pi s}\right) \\ \boldsymbol{B}_{n+1,n}^{\pi s} &= \boldsymbol{R}^{s\pi} \cdot \left((\boldsymbol{A}^{\pi})^{-1} \cdot \boldsymbol{B}_{n+1,n}^{\pi}\right) \end{aligned} \tag{13.51}$$

for solving only the soil temperature $\boldsymbol{T}_{n+1}^{s}$ at the new time plane $n+1$, where the modified matrix $\boldsymbol{A}^{s} - \boldsymbol{A}^{\pi s}$ represents the *Schur complement* [15]. Note that $\boldsymbol{A}^{\pi}$ is a local $(N_{\text{BHE}} \star \text{DOF}) \times (N_{\text{BHE}} \star \text{DOF})$ matrix, which is commonly not large $N_{\text{BHE}} \ll N_{\text{NP}}$ ($N_{\text{BHE}} < 1000$, DOF $= 8$ for 2U, DOF $= 4$ for 1U and DOF $= 3$ for CXA and CXC). Accordingly, the inverse $(\boldsymbol{A}^{\pi})^{-1}$ can be easily computed by a direct Gaussian matrix solution for each BHE. If $\boldsymbol{T}_{n+1}^{s}$ is solved from (13.50) the internal temperatures $\boldsymbol{T}_{n+1}^{\pi}$ for each exchanger can be simply recomputed from (13.49).

Using (13.50) and (13.49) a direct and non-sequential solution of complete temperature field for the soil and the BHE, $\boldsymbol{T}_{n+1}^{s}$, $\boldsymbol{T}_{n+1}^{\pi}$, appears possible. Basically, there is no need for an iterative solution of the coupled system (13.45), which is superior to the strictly iterative sequential strategy as used by Al-Khoury et al. [7, 8]. However, the condensed matrix system (13.50) with the Schur complement $\boldsymbol{A}^{s} - \boldsymbol{A}^{\pi s}$ has been shown frequently very stiff, particularly when the heat transfer coefficients dominate above thermal conduction and advection of the global system. In such cases numerical roundoff errors can distort the solution and balance errors occur in long-term or steady-state simulations. To prevent these harmful effects the solution of the severely ill-conditioned matrix system (13.50) is combined with an iterative correction strategy as follows:

$$\begin{aligned} &\text{starting solution } \tau = 0: && \begin{cases} (\boldsymbol{A}^{s} - \boldsymbol{A}^{\pi s}) \cdot \boldsymbol{T}_{n+1}^{s,\tau} = \boldsymbol{B}_{n+1,n}^{s} - \boldsymbol{B}_{n+1,n}^{\pi s} \\ \boldsymbol{T}_{n+1}^{\pi,\tau} = (\boldsymbol{A}^{\pi})^{-1} \cdot \left(\boldsymbol{B}_{n+1,n}^{\pi} - \boldsymbol{R}^{\pi s} \cdot \boldsymbol{T}_{n+1}^{s,\tau}\right) \end{cases} \\ &\text{iterative correction } \tau + 1: && \begin{cases} \boldsymbol{A}^{s} \cdot \boldsymbol{T}_{n+1}^{s,\tau+1} = \boldsymbol{B}_{n+1,n}^{s} - \boldsymbol{R}^{s\pi} \cdot \boldsymbol{T}_{n+1}^{\pi,\tau} \\ \boldsymbol{T}_{n+1}^{\pi,\tau+1} = (\boldsymbol{A}^{\pi})^{-1} \cdot \left(\boldsymbol{B}_{n+1,n}^{\pi} - \boldsymbol{R}^{\pi s} \cdot \boldsymbol{T}_{n+1}^{s,\tau+1}\right) \end{cases} \end{aligned} \tag{13.52}$$

where $\tau = 0, 1, 2, \ldots$ corresponds to an iteration counter. At each time plane we start with the Schur complement solution. It results the soil temperature $\boldsymbol{T}_{n+1}^{s,\tau}$ and the BHE temperature $\boldsymbol{T}_{n+1}^{\pi,\tau}$ at initial state $\tau = 0$. With known $\boldsymbol{T}_{n+1}^{\pi,\tau}$ the global soil matrix system (second row of matrix system (13.45)) is solved to find the new iterate for temperatures of soil $\boldsymbol{T}_{n+1}^{s,\tau+1}$ and accordingly of BHE $\boldsymbol{T}_{n+1}^{\pi,\tau+1}$. The iteration in (13.52) is repeated until a satisfactory convergence is achieved, such as

$$\|\boldsymbol{T}_{n+1}^{\pi,\tau+1} - \boldsymbol{T}_{n+1}^{\pi,\tau}\|_{L_p} \leq \epsilon \tag{13.53}$$

where ϵ is a given dimensionless error tolerance and $\|.\|_{L_p}$ denotes a suitable (maximum or RMS) error norm. Usually, only one iteration is required in transient simulations if the time step length Δt_n is chosen appropriately small. This is effectively controlled by using the adaptive time stepping strategy combined with predictor-corrector schemes as described above.

13.5.3 Suitable Meshing of BHE Nodes

In using either the analytical or numerical solution strategies a BHE is reduced to an internal well-type SPC occupied at a single node in a horizontal view on the 3D finite element mesh of the global problem. It appears similar to a well node, where a pumping well with a rate Q_w in the borehole is modeled at a singular node via a well function applied to the sink/source term for flow such that $Q_{hw} = -Q_w \delta(\boldsymbol{x} - \boldsymbol{x}_w)$, where $\delta()$ is the Dirac delta function and $\boldsymbol{x}_w$ are the well coordinates of the well node w.

Such types of nodal singularities in a mesh require specific considerations due to the following reasons. If inserting Q_w at a singular node w the resulting head value h_w in a flow simulation does not usually represent the head exactly at the physical borehole radius r_b; instead, the actually computed head h_w at the node w is to be deemed on a different radius, which is called *virtual radius* r_{virtual}; in regional models often larger than the real physical radius r_b. It can be shown that the virtual radius r_{virtual} is primarily dependent on the mesh discretization around the node w, represented by a nodal distance Δ (cf. Fig. 13.3). Accordingly, it has to be the goal in present modeling to design the mesh around those singular well nodes w in such a way that the virtual radius r_{virtual} meets at best the physical radius r_b of the well. In doing this, we introduce methods for tuning the mesh at BHE nodes [146].

A simple but efficient method represents the direct estimation of nodal distance Δ which follows the ideas by Nillert [391] developed for 2D horizontal regular meshes applied to wells in groundwater flow. Extending to conductive heat transport we find the following relationships, which are similar to potential flow. In a spatial discretization the conductive heat flux Q_{Tw} at the singular node w can be expressed by

$$Q_{Tw} = \vartheta \Phi_T (T_\Delta - T_{\text{virtual}}) \tag{13.54}$$

where T_Δ is the temperature at the distance Δ, T_{virtual} is the temperature at the virtual radius r_{virtual}, which must not be the physical BHE radius r_b, Φ_T is the heat transfer

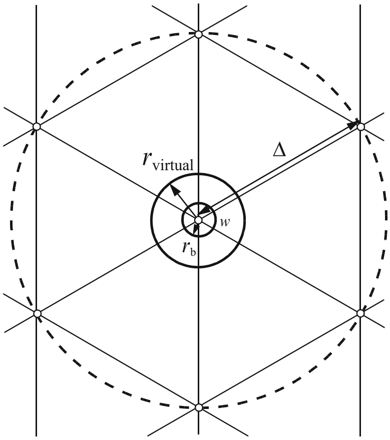


Fig. 13.3 Spatial discretization ($n = 6$) around a BHE 'well' node

coefficient and ϑ is a shape factor determined by the BHE-node surrounding mesh. For regular 2D meshes Nillert [391] derived:

$$\vartheta = n \tan\left(\tfrac{\pi}{n}\right) \tag{13.55}$$

where $n = 4, 5, 6, \ldots$ denotes the number of surrounding nodes, where $n = 6$ is typical for triangular horizontal meshes (see Fig. 13.3).

In contrast to the approximate solution (13.54), for a radially symmetric BHE we find the analytical (heat) well formula [33]

$$Q_{Tw}^{\text{ana}} = 2\pi \Phi_T \frac{T_\Delta - T_{\text{virtual}}}{\ln(\frac{\Delta}{r_{\text{virtual}}})} \tag{13.56}$$

Equating (13.56) and (13.54) it yields

$$\Delta = e^{\alpha} r_{\text{virtual}}\,, \qquad \alpha = \frac{2\pi}{n \tan\left(\frac{\pi}{n}\right)} \tag{13.57}$$

Equation (13.57) can be used to determine the required nodal distance Δ spacing from the BHE node if forcing the virtual radius to the borehole radius: $r_{\text{virtual}} = r_b$. It obtains for typical horizontal meshes

$$\Delta = a r_b\,, \qquad a = \begin{cases} 4.81 & \text{for} \quad n = 4 \\ 6.13 & \text{for} \quad n = 6 \\ 6.66 & \text{for} \quad n = 8 \end{cases} \tag{13.58}$$

Relation (13.58) represents a direct and effective estimation for an *optimal mesh* refinement around a BHE node. It will be shown further below (see Sect. 13.6.3) that those meshes which are designed by using criterion (13.58) can give optimal accuracy, even better than spatial discretizations over-refined $\Delta \leq r_b$ or coarse $\Delta > ar_b$ around BHE nodes.

An alternative method providing an iterative estimation of nodal distance Δ has been described in [146], which is applicable on the actual discretization. However, the effort of this iterative procedure can be high, particularly if applied to arrays of BHE. In practical applications, the above direct estimation method has shown often superior and sufficient. We note that all estimation methods assume that the heat transfer process is dominated by a radial conduction having no (or negligible) variation in the vertical direction.

13.6 Examples

13.6.1 Heat Transport for a Well Doublet System in a Layered Aquifer: 3D Modeling in Comparison to Analytical Solution

As an example of an open geothermal system let us consider the heat transport and exchange in a confined aquifer consisting of a high permeable layer overlain by a low permeable layer. A well doublet is installed in the high permeable layer, in which hot water is injected at the inlet well (source) and cool water is extracted from the outlet well (sink) as illustrated in Fig. 13.4. The distance between inlet well and outlet well is $2a$. It is assumed that heat advection dominates in the high permeable layer, while in the overlain low permeable layer only heat conduction is present. As a result, a heat plume establishes around the inlet well, which migrates toward the outlet well in time due to advection leading to a breakthrough of heated water in the outlet well at later time (Fig. 13.4). In this process heat is exchanged with the overlain low permeable layer due to the conductive heat transfer in vertical direction.

For the present problem an analytical solution can be found which is based on Muskat's dipolar potential flow [381]. Suppose a steady-state horizontal flow in the high permeable layer governed by the continuity and Darcy equations in the form, respectively,

$$\begin{gathered}\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} = 0 \\ q_x = -K\frac{\partial h}{\partial x}, \quad q_y = -K\frac{\partial h}{\partial y}\end{gathered} \tag{13.59}$$

neglecting density and viscosity effects and assuming isotropic conditions, the resulting potential equation for the hydraulic head h

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} = 0 \tag{13.60}$$

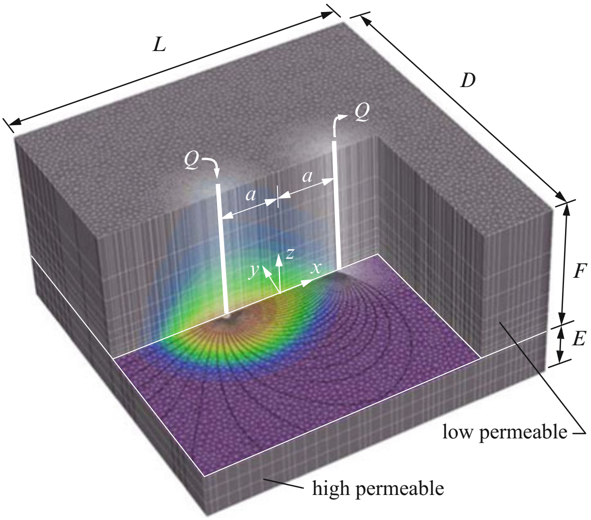


Fig. 13.4 Well doublet injecting hot water into a high permeable layer overlain by a low permeable layer. Advective-conductive heat plume distribution with pathlines in the lower layer and conductive heat exchange with the upper layer. Geometry and used mesh of model domain

gives with an imposed source at $Q(x, y) = Q(-a, 0)$ and an imposed sink at $Q(x, y) = Q(a, 0)$ of known discharge the following distribution within the $x - y-$plane for the hydraulic head

$$h = -\frac{Q}{8\pi KE} \ln\left(\frac{(x+a)^2 + y^2}{(x-a)^2 + y^2}\right) \tag{13.61}$$

and for the Darcy velocity components

$$\begin{aligned} q_x &= \frac{Q}{4\pi E}\left(\frac{x+a}{(x+a)^2 + y^2} - \frac{x-a}{(x-a)^2 + y^2}\right) \\ q_y &= \frac{Q}{4\pi E}\left(\frac{y}{(x+a)^2 + y^2} - \frac{y}{(x-a)^2 + y^2}\right) \end{aligned} \tag{13.62}$$

where E is the thickness of the high permeable layer (Fig. 13.4). The corresponding equations for the heat transport of the two-layer problem read

$$\begin{aligned} E\big(\varepsilon\rho c + (1-\varepsilon)\rho^s c^s\big)\tfrac{\partial T}{\partial t} + E\rho c(q_x \tfrac{\partial T}{\partial x} + q_y \tfrac{\partial T}{\partial y}) &= \Lambda^s \tfrac{\partial T}{\partial z} \qquad z \le 0 \quad \text{high permeable} \\ \rho^s c^s \tfrac{\partial T}{\partial t} &= \Lambda^s \tfrac{\partial^2 T}{\partial z^2} \qquad z > 0 \quad \text{low permeable} \end{aligned} \tag{13.63}$$

where it is assumed that pure advective transport governs in the high permeable layer (i.e., no heat conduction occurs in the $x - y-$plane of this layer, however, a vertical conductive heat exchange with the overlain layer exists), while advection is completely excluded in the low permeable layer. The solution of (13.63) for $T = T(x, y, z, t)$ subject to the IC and BC's

$$T(x,y,z,0) = T_0, \quad T(-a,0,0,t) = T_D, \quad \lim_{z\to\infty} T(x,y,z,t) = T_0 \tag{13.64}$$

can be obtained analytically [316, 467] leading to the following expression:

$$\frac{T(x,y,z,t)-T_0}{T_D-T_0} = \delta\big(t-I(x,y)\big)\ \mathrm{erfc}\frac{1}{2\sqrt{t-I(x,y)}}\left(\frac{\sqrt{\Lambda^s\rho^s c^s}}{E(\varepsilon\rho c+(1-\varepsilon)\rho^s c^s)}\right.$$
$$\left. I(x,y)+\sqrt{\frac{\rho^s c^s}{\Lambda^s}}\, z\right) \tag{13.65}$$

with

$$I(x,y) = \frac{4\pi E a^2 f}{Q}\left[1+2\cot\eta\ \arctan\left(\frac{\tan\eta/2\ (\tanh\xi/2-1)}{1+\tanh\xi/2\ \tan^2\eta/2}\right)-\right.$$
$$\left.\frac{\sinh\xi}{\cosh\xi+\cos\eta}\right]\sin^{-2}\eta \qquad \text{for} \quad |\cos\eta|\neq 1 \tag{13.66}$$

$$I(x,y) = \frac{4\pi E a^2 f}{3Q}\left[1-\frac{\sinh\xi}{\cosh\xi+\cos\eta}\left(1+\frac{\cos\eta}{\cosh\xi+\cos\eta}\right)\right]$$
$$\text{for} \quad |\cos\eta| = 1 \tag{13.67}$$

and

$$\xi = \tfrac{1}{2}\ln\left(\frac{(x-a)^2+y^2}{(x+a)^2+y^2}\right), \quad \eta = \arctan\frac{2ay}{a^2-x^2-y^2}, \quad f = \frac{\varepsilon\rho c}{\varepsilon\rho c+(1-\varepsilon)\rho^s c^s} \tag{13.68}$$

where $\delta()$ is the Dirac delta function.

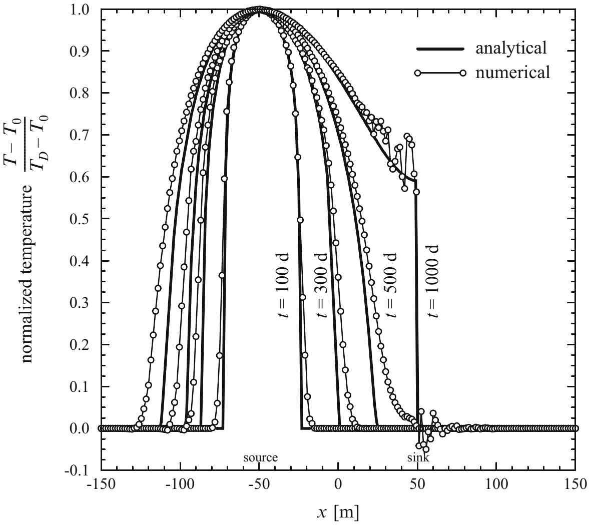
For the numerical simulations a 3D finite element mesh of moderate resolution is used as shown in Fig. 13.4. It consists of 215,670 pentahedral prismatic elements with 116,970 nodes in total, formed by 15 slices in the vertical z–direction. The mesh is suitably refined in the $x-y$–plane around the doublet wells (smallest element is about 0.7 m there) while the size of elements gradually increases with the distance from the wells. In the vertical direction the low permeable layer with a total thickness F is subdivided by ten sublayers and the high permeable layer with a total thickness E by four sublayers of variable element thicknesses. At the contact zone ($z=0$) between the low and high permeable layer the element thickness is chosen 0.1 m. With the vertical distance from the contact zone the element thickness gradually increases. The parameters and conditions used in the numerical simulation are summarized in Table 13.1. Unspecified BC's represent boundaries, at which natural BC's are imposed. To compare the numerical simulation with the analytical solution the chosen domain should be sufficiently large so that the adiabatic (no-heat flux) condition $-(\boldsymbol{\Lambda}\cdot\nabla T)\cdot\boldsymbol{n}=0$ imposed on the outer enclosing boundaries can

Table 13.1 Parameters and conditions used for the well doublet problem

Quantity	Symbol	Value	Unit
Domain shown in Fig. 13.4			
Domain length	L	300	m
Domain width	D	300	m
Thickness of low permeable layer	F	30	m
Thickness of high permeable layer	E	10	m
Specific storage coefficient	S_o	0	m^{-1}
Longitudinal dispersivity	β_L	0	m
Transverse dispersivity	β_T	0	m
Constant doublet discharge	Q	150	$\text{m}^3\,\text{d}^{-1}$
Low permeable layer			
Isotropic hydraulic conductivity	K	10^{-8}	$\text{m}\,\text{s}^{-1}$
Porosity	ε	0	1
Volumetric heat capacity of solid	$\rho^s c^s$	$2.52 \cdot 10^6$	$\text{J}\,\text{m}^{-3}\,\text{K}^{-1}$
Heat conductivity of solid	Λ^s	3	$\text{J}\,\text{m}^{-1}\,\text{s}^{-1}\,\text{K}^{-1}$
Anisotropy factor	$\Xi_{\text{aniso}}^{\Lambda}$	1	1
High permeable layer			
Isotropic hydraulic conductivity	K	10^{-4}	$\text{m}\,\text{s}^{-1}$
Porosity	ε	0.3	1
Volumetric heat capacity of fluid	ρc	$4.2 \cdot 10^6$	$\text{J}\,\text{m}^{-3}\,\text{K}^{-1}$
Volumetric heat capacity of solid	$\rho^s c^s$	$2.52 \cdot 10^6$	$\text{J}\,\text{m}^{-3}\,\text{K}^{-1}$
Heat conductivity of fluid	Λ	10^{-6}	$\text{J}\,\text{m}^{-1}\,\text{s}^{-1}\,\text{K}^{-1}$
Heat conductivity of solid	Λ^s	10^{-6}	$\text{J}\,\text{m}^{-1}\,\text{s}^{-1}\,\text{K}^{-1}$
Anisotropy factor	$\Xi_{\text{aniso}}^{\Lambda}$	10^9	1
IC and BC's			
Initial condition (IC) of T (13.6)	T_0	10	°C
Multilayer well BC of inlet at $(x = -a, y = 0, -E \leq z \leq 0)$	$Q_w = -Q$	−150	$\text{m}^3\,\text{d}^{-1}$
Multilayer well BC of outlet at $(x = a, y = 0, -E \leq z \leq 0)$	$Q_w = Q$	150	$\text{m}^3\,\text{d}^{-1}$
Dirichlet-type BC for h on outer boundary at $(-\frac{L}{2} \leq x \leq \frac{L}{2}, y = \pm\frac{D}{2}, -E \leq z \leq 0)$ and $(x = \pm\frac{L}{2}, -\frac{D}{2} \leq y \leq \frac{D}{2}, -E \leq z \leq 0)$	h_D	(13.61)	m
Dirichlet-type BC for T on inlet well at $(x = -a, y = 0, -E \leq z \leq 0)$	T_D	60	°C
FEM			
Unstructured 3D mesh of 215,670 linear pentahedra, GFEM (no upwind)			
Initial time step size[a]	Δt_0	10^{-3}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Simulation time period	t_{end}	$1,000$	d

[a] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$ and maximum time step size $\Delta t_{\max} = 3\,\text{d}$

Fig. 13.5 Simulated versus analytical temperature profiles $\frac{T-T_0}{T_D-T_0}$ along $x-$axis at $y = z = 0$ and different times t in days. Simulation results obtained by using GFEM and GLS 1st-order accurate FE/BE predictor-corrector time integration



be maintained. To model exactly the potential flow field in the high permeable layer, we impose the known hydraulic head values h from (13.61) as Dirichlet-type BC's on the outer enclosing boundaries.

A difficulty in comparison with the analytical solution (13.65) is in the pure advection of heat in horizontal direction of the high permeable layer, while heat conduction in horizontal direction is suppressed. To obtain stable and sufficiently accurate numerical results, GFEM without any spurious upwinding, however, with the stabilized GLS 1st-order accurate FE/BE predictor-corrector time integration is preferred. (Note that the use of a 2nd-order accurate AB/TR scheme would not give stable solutions in the present case.) Additionally, a very small amount of heat conduction (however, without thermodispersion) in horizontal direction of the high permeable layer is admitted, while in the vertical direction of the high permeable layer heat conduction can be taken suitably large (enforced by a properly large anisotropy factor, cf. Table 13.1).

For the simulation of the heat transport over a time period of 1,000 days only 353 implicit time steps are required. The achieved computational results are compared with the analytical solution (13.65) in Figs. 13.5 and 13.6. The agreements are reasonably well. As revealed in Fig. 13.5 at early times the results compare better to the analytical findings, however, with elapsing time the numerical solutions become more diffused due to numerical dispersion effect caused by the 1st-order accurate FE/BE scheme. At times when the heat plume reaches the outlet local oscillations in the temperature profile can be observed (Fig. 13.5). Then, sharp temperature profiles begin to be established at the outlet well, where heated and cool water joins at the single well-type SPC node(s). Note that the situation can be improved if the well-type SPC condition of the outlet well is replaced by a Neumann-type flux BC imposed on a fully discretized borehole geometry of the outlet.

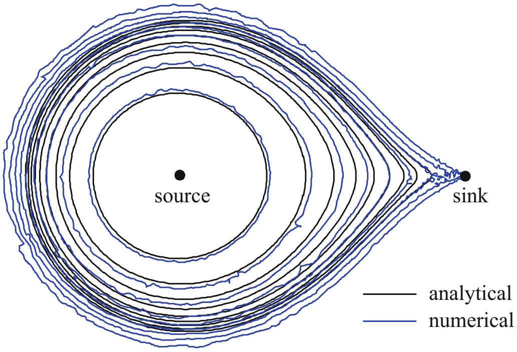


Fig. 13.6 Simulated temperature contours $\frac{T-T_0}{T_D-T_0}$ in the $x-y-$plane at $z=0$ and $t=600$ d in comparison to the analytical distribution. Used contouring interval of normalized temperature is 0.1

13.6.2 Numerical vs. Analytical Solutions of BHE for Steady-State Conditions and Given Temperature at Borehole Wall

We directly compare the numerical and analytical solution strategies by Al-Khoury et al. [6–8] and Eskilson and Claesson [159] for local BHE problems under steady-state conditions. The analytical BHE solutions are compared to the numerical BHE results for CXA, CXC, 1U and 2U-type BHE configurations with the parameters as listed in Tables 13.2–13.4. Since the soil temperature T_s is here specified as a BC the properties of soil become irrelevant for the present comparison analysis. The thermal resistances and heat transfer coefficients as summarized in Table 13.4 are computed from the formula given in Sects. M.2 and M.3, respectively, of Appendix M. In the simulation models only the inner borehole is discretized, where BC's for the solid temperature T_s are prescribed at the BHE node patch as exhibited in Fig. 13.7. For the vertical discretization 100 layers are used, i.e., $\Delta z = 1$ m.

The numerical results versus the analytical solutions in form of steady-state vertical temperature profiles of pipe(s)-in, pipe(s)-out and grout zone(s) are shown in Fig. 13.8 for each of the CXA, CXC, 1U and 2U-type BHE configurations. As evidenced in all cases the agreement is nearly perfect.

13.6.3 Transient Solution of Coaxial BHE System

We consider a BHE coaxial pipe system of annular inlet (CXA type) with parameters as listed in Tables 13.5 and 13.6. The soil domain measures 100×100 m in horizontal directions and 100 m in depth. It is assumed that the soil is impervious and no groundwater exists, i.e., $\varepsilon = 0$ and $\boldsymbol{q} = \mathbf{0}$. The used mesh for the BHE solution is shown in Fig. 13.9. The BHE is located in the center of the domain, where the mesh is locally refined. For the vertical discretization 100 layers are applied. Two variants of heat injections are considered. The first one refers to a small-rate injection with

Table 13.2 BHE parameters used for analytical comparison

Quantity	Symbol	Value	Unit
Depth of borehole	L	100	m
Borehole diameter	D	10	cm
Pipe measures listed in Table 13.3			
Reference temperature	T_0	10	°C
Thermal conductivities of pipe walls	Λ^{π}_{i1}, Λ^{π}_{o1}	0.38	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Total flow discharge of refrigerant	Q_r	21.86	$\mathrm{m^3\,d^{-1}}$
Total heat input rate	$\lvert Q_{Tw}\rvert$	$6.3216 \cdot 10^9$	$\mathrm{J\,d^{-1}}$
Volumetric heat capacity of refrigerant	$\rho^r c^r$	$4.1312 \cdot 10^6$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of refrigerant	Λ^r	0.6405	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Dynamic viscosity of refrigerant	μ^r	$0.54741 \cdot 10^{-3}$	$\mathrm{kg\,m^{-1}\,s^{-1}}$
Mass density of refrigerant	ρ^r	$0.9881 \cdot 10^3$	$\mathrm{kg\,m^{-3}}$
Volumetric heat capacity of grout	$\rho^g c^g$	$2.19 \cdot 10^6$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of grout	Λ^g	2.3	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Thermal resistances and heat transfer coefficients listed in Table 13.4			
BC's			
Dirichlet-type BC for soil T on outer surface	$T_D = T_s$	10	°C
Dirichlet-type BC of inlet T at pipe(s)-in[a]	$T_D = T_i$	80	°C
FEM			
3D cylindric mesh for the inner borehole consisting of 100 layers with 17,200 pentahedra			

[a] $T_i = \lvert Q_{Tw}\rvert/(\rho^r c^r Q_r) + T_0$

Table 13.3 Pipe measures for the BHE configurations[a]

Parameter	Symbol	CXA	CXC	1U	2U	Unit
Outer diameter of pipe-in[b]	d^o_{i1}	5	2.4	3.2	3.2	cm
Outer diameter of pipe-out[c]	d^o_{o1}	2.4	5	3.2	3.2	cm
Pipe-in wall thickness[d]	b_{i1}	4	3	2.9	2.9	mm
Pipe-out wall thickness[e]	b_{o1}	3	4	2.9	2.9	mm
Pipe distance	w	–	–	6	4.242	cm

[a] See definitions shown in Figs. M.1 and M.2 of Appendix M

[b] $d^o_{i1} = 2r^o_{i1} = 2r^o_{i2}$

[c] $d^o_{o1} = 2r^o_{o1} = 2r^o_{o2}$

[d] $b_{i1} = r^o_{i1} - r^i_{i1} = r^o_{i2} - r^i_{i2}$

[e] $b_{o1} = r^o_{o1} - r^i_{o1} = r^o_{o2} - r^i_{o2}$

laminar flow in the coaxial pipes, which is highly driven by thermal conduction. On the other hand, a turbulent flow regime is applied, where advective heat transport in the pipe system is more apparent. In both variants in the time range ($0 < t \leq 90\,\mathrm{d}$) water with a temperature of 80 °C is injected at the annular pipe-in. At later times ($90\,\mathrm{d} < t \leq 180\,\mathrm{d}$) the injection temperature amounts to 10 °C.

Both the Al-Khoury et al.'s numerical BHE method and the Eskilson and Claesson's analytical BHE method are applied. The present FEFLOW results are compared to a fully discretized FDM solution for an axisymmetric 2D formulation of the problem as given by Heidemann [234]. Heidemann has discretized the

Table 13.4 Thermal resistances R and heat transfer coefficients Φ for the BHE configurations

Parameter	Symbol	CXA	CXC	1U	2U	Unit
Thermal resistances:						
Pipe-in to grout	R_{fig}	0.10874	–	0.15577	0.14485	$\mathrm{m\,s\,K\,J^{-1}}$
Pipe-in to pipe-out	R_{ff}	0.13037	0.13037	–	–	$\mathrm{m\,s\,K\,J^{-1}}$
Pipe-out to grout	R_{fog}	–	0.10874	0.15577	0.14485	$\mathrm{m\,s\,K\,J^{-1}}$
Grout to grout	R_{gg}	–	–	0.11516	–	$\mathrm{m\,s\,K\,J^{-1}}$
Grout to grout 1	R_{gg1}	–	–	–	0.00031	$\mathrm{m\,s\,K\,J^{-1}}$
Grout to grout 2	R_{gg2}	–	–	–	0.11776	$\mathrm{m\,s\,K\,J^{-1}}$
Grout to soil	R_{gs}	0.01626	0.01626	0.02574	0.06833	$\mathrm{m\,s\,K\,J^{-1}}$
Heat transfer coefficients:						
Pipe-in to grout	Φ_{fig}	69.698	–	77.993	83.877	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Pipe-in to pipe-out	Φ_{ff}	135.64	135.64	–	–	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Pipe-out to grout	Φ_{fog}	–	69.698	77.993	83.877	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to grout	Φ_{gg}	–	–	66.796	–	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to grout 1	Φ_{gg1}	–	–	–	48, 489	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to grout 2	Φ_{gg2}	–	–	–	65.323	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to soil	Φ_{gs}	195.74	195.74	190.24	143.32	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$

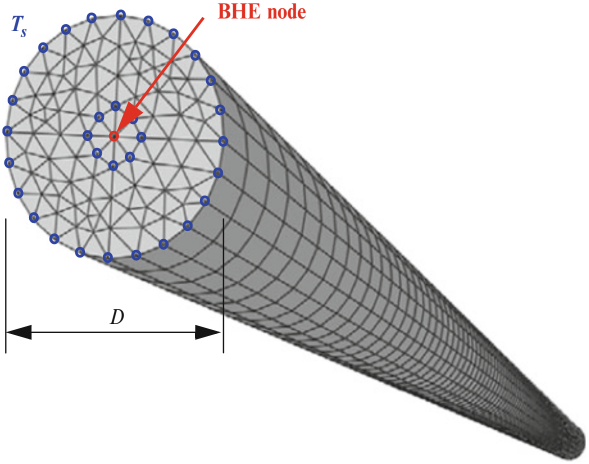


Fig. 13.7 Discretized inner borehole with temperature BC of soil T_s (indicated on *top slice*)

meridional cross section by a 72 × 113 grid. The radial extension is taken with 50 m. His grid has been gradually spaced along the radial direction ranging from 1.5 mm up to 1 m. Heidemann used variable time steps between 30 min and 4 h.

The outlet temperature histories computed by the numerical and analytical BHE methods in comparison to Heidemann's solution are displayed in Fig. 13.10 for the laminar flow and in Fig. 13.11 for the turbulent flow. The results are in a reasonable agreement. For the turbulent case we find an excellent agreement between Heidemann's and the analytical BHE solution as evidenced in Fig. 13.11a.

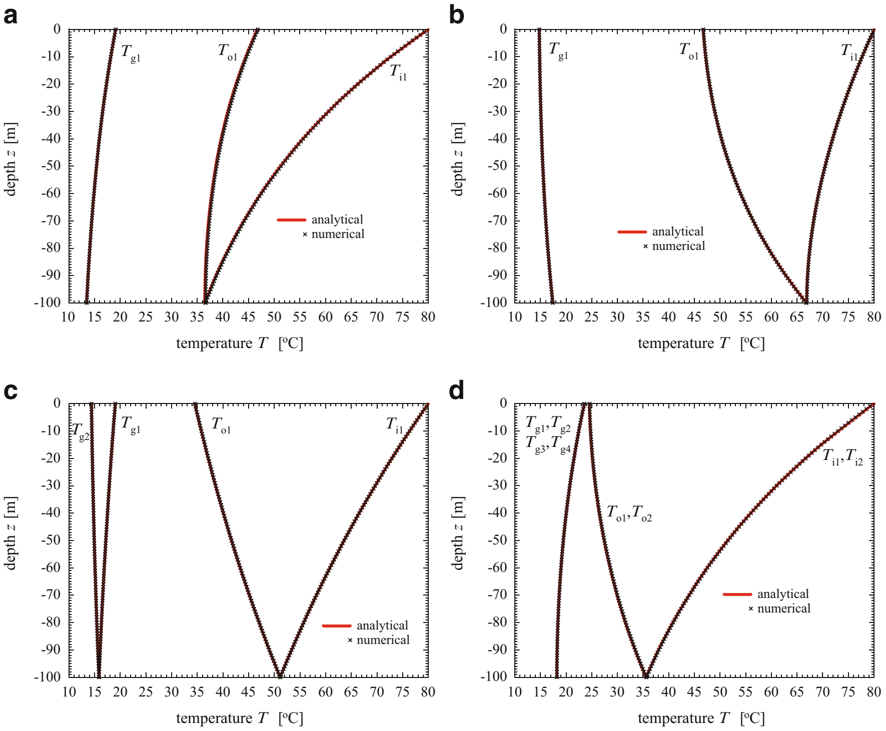


Fig. 13.8 Analytical vs. numerical temperature distribution for (**a**) CXA, (**b**) CXC, (**c**) 1U and (**d**) 2U BHE configuration

We have to note that the present analytical BHE solutions are invalid for variations in a time scale shorter than about 3.5 h according to the limit (M.78). Using limit (M.79) input variations cannot be simulated even below about 10 h for laminar flow and about 4 h for turbulent flow. In Figs. 13.10b and 13.11b the short-term temperature behavior of the analytical and numerical BHE methods are shown for the laminar and turbulent flow cases, respectively. They reveal how the analytical method overestimates the outlet temperature at transient input situations. However, these errors vanishes in long-term predictions if no longer input variations occur as depicted in Figs. 13.10a and 13.11a. It has been shown necessary to assign a high thermal conductivity Λ^s with an anisotropic factor $\Xi^{\Lambda}_{\text{aniso}}$ for the inner BHE surplus according to Table 13.5.

The present turbulent flow case of a single CXA-type BHE gives opportunity for a mesh convergence study, where the level of mesh refinement around the singular BHE node is systematically increased. This will reflect the statements of Sect. 13.5.3 regarding an optimal mesh design for BHE solutions. We test the accuracy of the solution for a stepwise local refinement of mesh Υ around the BHE node according to (cf. Fig. 13.12)

Table 13.5 Parameters of the CXA exchanger problem

Quantity	Symbol	Value	Unit
Depth of borehole	L	100	m
Borehole diameter	D	10	cm
Outer diameter of pipe-in	d_{i1}^{o}	5	cm
Outer diameter of pipe-out	d_{o1}^{o}	2.4	cm
Pipe-in wall thickness	b_{i1}	4	mm
Pipe-out wall thickness	b_{o1}	3	mm
Reference temperature	T_0	10	°C
Thermal conductivities of pipe walls	$\Lambda_{i1}^{\pi}, \Lambda_{o1}^{\pi}$	0.38	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Total flow discharge of refrigerant:			
Laminar flow	Q_r^{laminar}	1.0931	$\mathrm{m^3\,d^{-1}}$
Turbulent flow	$Q_r^{\text{turbulent}}$	21.8624	$\mathrm{m^3\,d^{-1}}$
Total heat input rate:			
Laminar flow	$\lvert Q_{Tw}^{\text{laminar}}(t)\rvert$	$\begin{cases} 3.1602 \cdot 10^{8} & (0 < t \leq 90\,\mathrm{d}) \\ 0 & (90\,\mathrm{d} < t \leq 180\,\mathrm{d}) \end{cases}$	$\mathrm{J\,d^{-1}}$
Turbulent flow	$\lvert Q_{Tw}^{\text{turbulent}}(t)\rvert$	$\begin{cases} 6.3203 \cdot 10^{9} & (0 < t \leq 90\,\mathrm{d}) \\ 0 & (90\,\mathrm{d} < t \leq 180\,\mathrm{d}) \end{cases}$	$\mathrm{J\,d^{-1}}$
Volumetric heat capacity of refrigerant	$\rho^r c^r$	$4.13 \cdot 10^{6}$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of refrigerant	Λ^r	0.48	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Dynamic viscosity of refrigerant	μ^r	$0.52 \cdot 10^{-3}$	$\mathrm{kg\,m^{-1}\,s^{-1}}$
Mass density of refrigerant	ρ^r	$0.988 \cdot 10^{3}$	$\mathrm{kg\,m^{-3}}$
Volumetric heat capacity of grout	$\rho^g c^g$	$2.19 \cdot 10^{6}$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of grout	Λ^g	2.3	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Porosity of soil	ε	0	1
Volumetric heat capacity of soil	$\rho^s c^s$	$2.21 \cdot 10^{6}$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of soil	Λ^s	2.2	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Anisotropy factor of soil	$\Xi_{\text{aniso}}^{\Lambda}$	1	1
Thermal conductivity of BHE surplus	Λ^s	10^{3}	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Anisotropy factor of BHE surplus	$\Xi_{\text{aniso}}^{\Lambda}$	0	1
Thermal resistances and heat transfer coefficients listed in Table 13.6			
IC and BC			
Initial condition (IC) of T (13.6)	T_0	10	°C
Dirichlet-type BC for T at pipe-in	$T_D = T_i(t)$[a]	Variable	°C
FEM			
Nonuniform 3D mesh consisting of 100 layers with 239,100 pentahedra shown in Fig. 13.9			
Vertical space increment	Δz	1	m
Initial time step size[b]	Δt_0	10^{-6}	d
RMS error tolerance (FE/BE)	ϵ	10^{-3}	1
Simulation time period	t_{end}	180	d

[a] $T_i(t) = |Q_{Tw}(t)|/(\rho^r c^r Q_r) + T_0$

[b] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$

Table 13.6 Thermal resistances R and heat transfer coefficients Φ for the CXA exchanger problem

	Thermal resistance			Heat transfer		
Parameter	Symbol	Value	Unit	Symbol	Value	Unit
Laminar flow						
Pipe-in to grout	R_{fig}	0.14312	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{fig}	52.955	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Pipe-in to pipe-out	R_{ff}	0.33963	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{ff}	52.068	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to soil	R_{gs}	0.01626	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{gs}	195.74	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Turbulent flow						
Pipe-in to grout	R_{fig}	0.10932	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{fig}	69.326	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Pipe-in to pipe-out	R_{ff}	0.13183	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{ff}	134.14	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to soil	R_{gs}	0.01626	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{gs}	195.74	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$

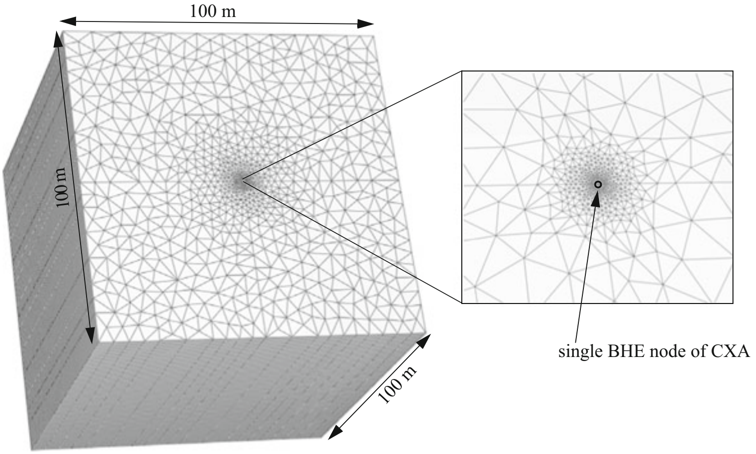


Fig. 13.9 Finite element mesh used for CXA-type BHE model consisting of 239,100 pentahedral prisms. Vertical discretization concerns 100 layers

$$\Upsilon_\ell \quad \ell = 0, 1, 2, \ldots, 8 \tag{13.69}$$

where ℓ is the refinement level of mesh Υ_ℓ. Starting with Υ_0 consisting of a regular triangular tessellation characterized by a BHE nodal distance Δ of about 4.42 m, the number of triangular prisms N_E and total number of nodes N_P then increase according to the refinement level ℓ, while the BHE nodal distance Δ is halved in value for each refinement level ℓ:

$$\begin{aligned} N_E &= 32(32+\ell)\cdot(N_S-1) \\ N_P &= [16(34+\ell)+1]\cdot N_S \\ \Delta_\ell &= 2^{-\ell}\Delta, \quad \Delta \approx \tfrac{L}{32}\sqrt{2} = 4.42\,\mathrm{m}, \quad N_S = 21 \end{aligned} \tag{13.70}$$

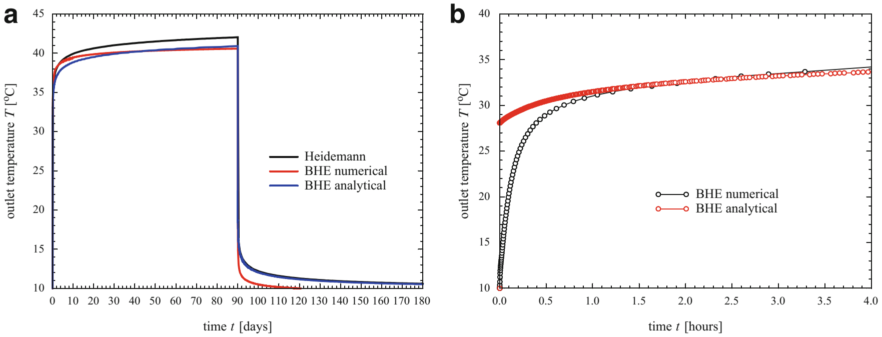


Fig. 13.10 (**a**) Long-term and (**b**) short-term temperature history at pipe outlet of the CXA-type BHE for laminar flow. Long-term temperature history is compared to Heidemann's solution

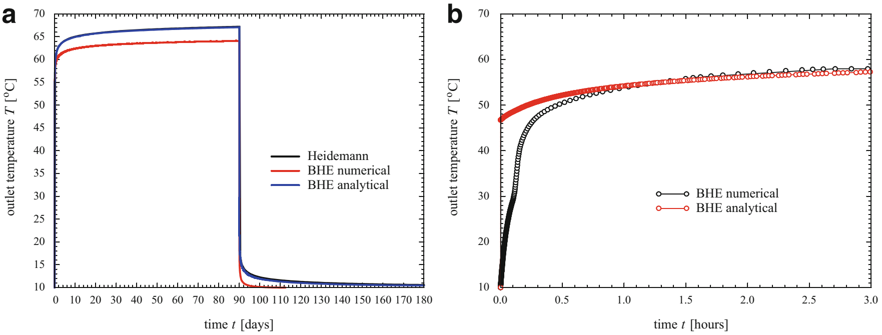


Fig. 13.11 (**a**) Long-term and (**b**) short-term temperature history at pipe outlet of the CXA-type BHE for turbulent flow. Long-term temperature history is compared to Heidemann's solution

where the BHE node (in the central position of the domain) is locally refined from level to level ℓ (see Fig. 13.12). For the mesh convergence test only a vertical discretization consisting of 20 layers (number of slices $N_S = 21$) with a vertical spacing of $\Delta z = L/(N_S - 1) = 5\,\text{m}$ is considered.

The simulations by using the analytical BHE method are performed up to a maximum refinement level of $\ell = 8$. At that level the BHE nodal distance with about 1.7 cm is clearly smaller than the physical borehole radius of $r_b = D/2 = 5\,\text{cm}$. Using the estimation (13.58) from Sect. 13.5.3 we can expect an optimal BHE nodal distance Δ of about 0.333 m (with $n = 8$), which would require a refinement level ℓ of about 4 ($\Delta_4 \approx 0.276\,\text{m}$) to attain suited accuracy. Indeed, the simulations reveal that the best agreement to Heidemann's reference solution is for Υ_4 as evidenced in Fig. 13.13 for the turbulent flow case. As revealed both coarse meshes ($\Upsilon_\ell,\ \ell < 4$) and higher dense meshes ($\Upsilon_\ell,\ \ell > 4$) under- and overestimates, respectively, the reference solution for the outlet temperature. If the nodal distance falls below the physical borehole radius r_b the elements within $\Delta \leq r_b$ have to assigned to a high thermal conductivity to break the further increase of the temperature at the borehole.

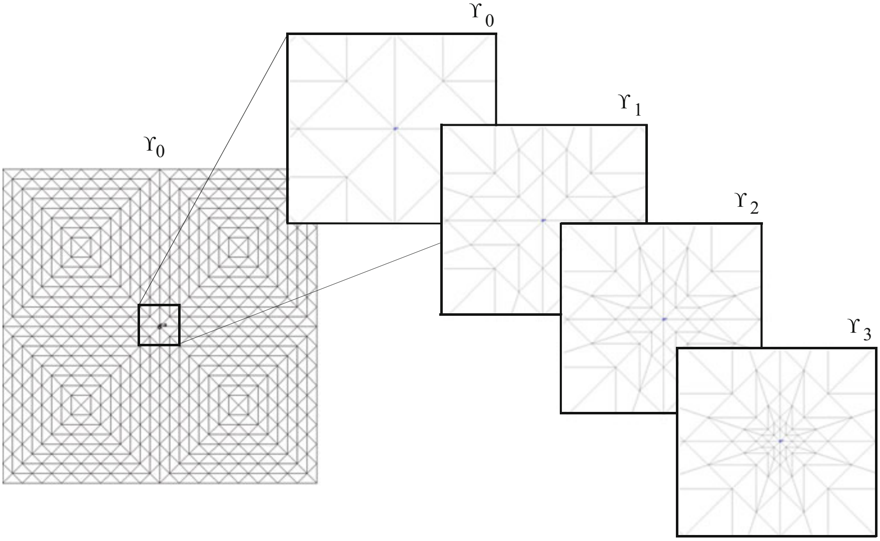


Fig. 13.12 Different mesh refinement levels for CXA-type BHE located in the center of the domain. Vertical discretization concerns 20 layers

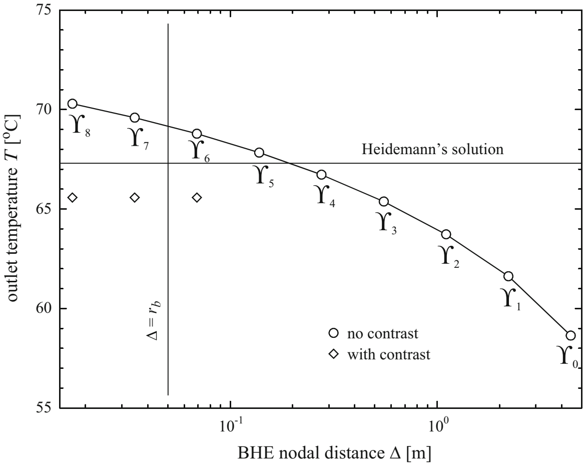


Fig. 13.13 Outlet temperature at $t = 90\,\mathrm{d}$ of the CXA-type BHE for turbulent flow versus the BHE nodal distance Δ. Refinement levels Υ_ℓ ($\ell = 0, \ldots, 8$) in comparison to Heidemann's reference solution. For levels $\ell = 6 - 8$ solutions with high contrast of the thermal conductivity $\Lambda^s = 10^3\,\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$ for elements smaller than physical borehole radius $r_b = 0.05\,\mathrm{m}$ are also incorporated. Analytical BHE method is used

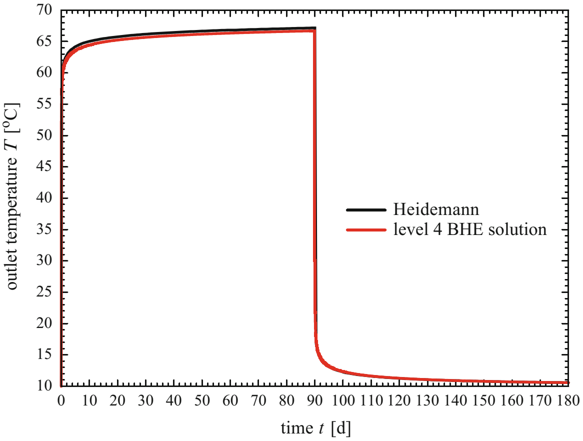


Fig. 13.14 Outlet temperature history of the CXA-type BHE for turbulent flow simulated with optimal mesh of refinement level 4, Υ_4 (analytical BHE solution) compared to Heidemann's solution

The results for the optimal Υ_4 mesh give very good agreement with Heidemann's reference solution as shown in Fig. 13.14 for the full history of outlet temperature. Although the mesh of level Υ_4 is about ten times coarser (consisting only of 23,040 pentahedral elements) than the mesh studied above (Fig. 13.9) consisting of 239,100 pentahedral elements, the quality of the results is comparable (cf. Figs. 13.14 vs. 13.11a).

13.6.4 BHE Solution Versus Fully Discretized 3D Model (FD3DM) Solution Applied to a 2U Exchanger

Comparisons between the proposed BHE solution and a fully discretized 3D model solution (FD3DM) are performed for heating operation of a 2U configuration located in central position of a confined aquifer domain measuring 20 × 20 m in horizontal directions and 55 m in depth. The used meshes for both solutions are shown in Fig. 13.15 revealing a much more refined tessellation for FD3DM to discretize appropriately the interior geometric structure of the 2U exchanger. In both meshes, however, the vertical discretization is the same by using 55 layers. For the 2U exchanger problem the used parameters and conditions are summarized in Tables 13.7 and 13.8. A steady-state groundwater flow with a constant head gradient between the left and right boundary of 10^{-3} is assumed. Unspecified BC's represent boundaries, at which natural BC's are imposed, i.e., no-flow boundary and adiabatic (no-heat flux) boundary. In FD3DM 1D discrete feature (fracture) elements have been used to model the internal pipes. It was necessary to assign the inner pipe surplus to a high thermal conductivity of solid with anisotropy. For the surplus we took a value of $\Lambda^s = 10^3\,\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$ with an anisotropy factor of $\Xi^{\Lambda}_{\text{aniso}} = 0$. In the surplus we use a porosity ε of zero.

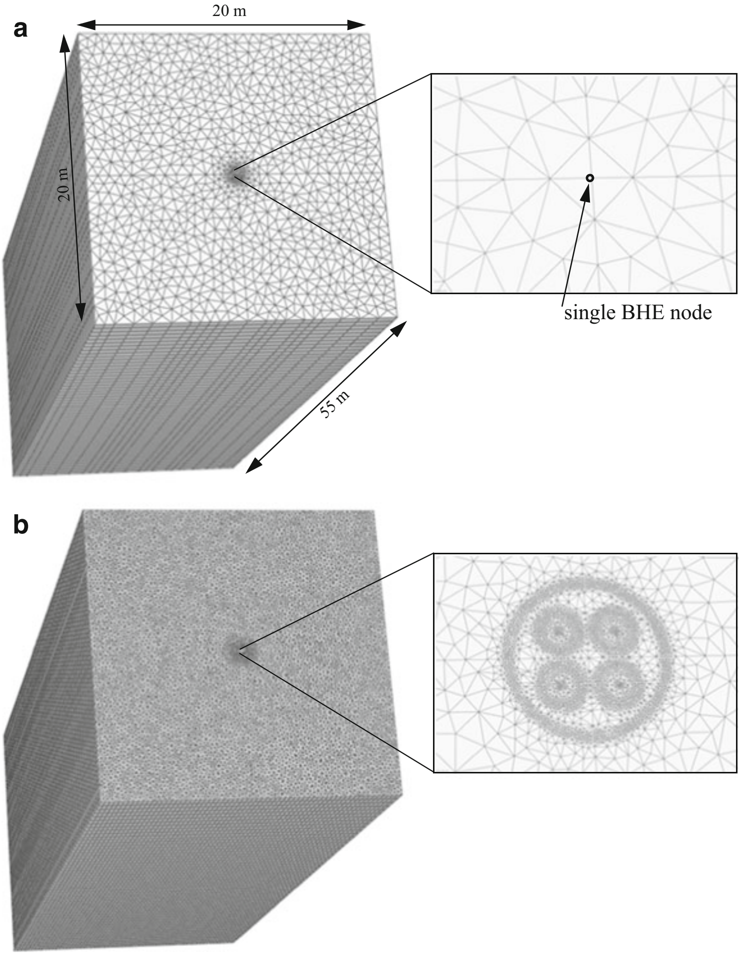


Fig. 13.15 Finite element meshes for (**a**) BHE consisting of 130,185 pentahedral elements and (**b**) FD3DM consisting of 1,204,665 pentahedral elements. Both meshes are vertically discretized by 55 numerical layers

A comparison between the BHE solutions to the FD3DM is shown in Fig. 13.16 for the short-term outlet temperature history, in Fig. 13.17 for the long-time outlet temperature history and in Fig. 13.18 for the vertical temperature profile after 12 h. As revealed the agreement between the different solutions is quite well. For long-term predictions the analytical BHE simulation has shown reasonably accurate and fast, while the numerical BHE computations became superior to the analytical BHE solution at short-term predictions and in a well agreement with the FD3DM results from beginning. In Fig. 13.18 the vertical temperature profile of grout is not evaluated for FD3DM because the grout temperature considerably varies within the mesh nodes in the borehole at that early time.

Table 13.7 Parameters and conditions used for the 2U exchanger problem

Quantity	Symbol	Value	Unit
Depth of borehole	L	55	m
Borehole diameter	D	12	cm
Outer diameter of pipes-in/pipes-out	d_{i1}^o, d_{i2}^o, d_{o1}^o, d_{o2}^o	3.2	cm
Pipes-in/pipes-out wall thickness	b_{i1}, b_{i2}, b_{o1}, b_{o2}	2.9	mm
Pipe distance	w	4.2	cm
Reference temperature	T_0	10	°C
Thermal conductivities of pipe walls	Λ_{i1}^π, Λ_{i2}^π, Λ_{o1}^π, Λ_{o2}^π	0.38	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Total flow discharge of refrigerant	Q_r	38.284	$\mathrm{m^3\,d^{-1}}$
Total heat input rate	$\lvert Q_{Tw}\rvert$	$6.3242 \cdot 10^9$	$\mathrm{J\,d^{-1}}$
Volumetric heat capacity of refrigerant	$\rho^r c^r$	$4.13 \cdot 10^6$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of refrigerant	Λ^r	0.65	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Dynamic viscosity of refrigerant	μ^r	$0.52 \cdot 10^{-3}$	$\mathrm{kg\,m^{-1}\,s^{-1}}$
Mass density of refrigerant	ρ^r	$0.938 \cdot 10^3$	$\mathrm{kg\,m^{-3}}$
Volumetric heat capacity of grout	$\rho^g c^g$	$2.19 \cdot 10^6$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of grout	Λ^g	2.3	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Porosity of soil	ε	0.2	1
Porosity of surplus	ε	0	1
Volumetric heat capacity of groundwater	ρc	$4.2 \cdot 10^6$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Volumetric heat capacity of soil	$\rho^s c^s$	$2.405 \cdot 10^6$	$\mathrm{J\,m^{-3}\,K^{-1}}$
Thermal conductivity of groundwater	Λ	0.65	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Thermal conductivity of soil	Λ^s	2.46	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Anisotropy factor of soil	$\Xi_{\text{aniso}}^{\Lambda}$	1	1
Thermal conductivity of BHE surplus	Λ^s	10^3	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$
Anisotropy factor of BHE surplus	$\Xi_{\text{aniso}}^{\Lambda}$	0	1
Longitudinal thermodispersivity of aquifer	β_L	0.5	m
Transverse thermodispersivity of aquifer	β_T	0.05	m
Thermal resistances and heat transfer coefficients listed in Table 13.8			
Flow BC's			
Dirichlet-type BC for h at left boundary (all slices)	h_D	0	m
Dirichlet-type BC for h at right boundary (all slices)	h_D	−0.02	m
Heat IC and BC's			
Initial condition (IC) of T (13.6)	T_0	10	°C
Dirichlet-type BC for T at pipe-in	$T_D = T_i^{\text{a}}$	50	°C
FEM			
Nonuniform 3D meshes consisting of 55 numerical layers as shown in Fig. 13.15, GFEM			
Vertical space increment	Δz	1	m

(continued)

Table 13.7 (continued)

Quantity	Symbol	Value	Unit
BHE solutions			
Initial time step size[b]	Δt_0	10^{-8}	d
RMS error tolerance (FE/BE)	ϵ	10^{-3}	1
FD3DM solution			
Initial time step size[c]	Δt_0	10^{-6}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	365	d

[a] $T_i = |Q_{Tw}|/(\rho^r c^r Q_r) + T_0$

[b] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$

[c] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 5$

Table 13.8 Thermal resistances R and heat transfer coefficients Φ for the 2U exchanger problem

	Thermal resistance			Heat transfer		
Parameter	Symbol	Value	Unit	Symbol	Value	Unit
Pipe-in to grout	R_{fig}	0.1326	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{fig}	91.624	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Pipe-out to grout	R_{fog}	0.1326	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{fog}	91.624	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to grout 1	R_{gg1}	0.02077	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{gg1}	802.43	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to grout 2	R_{gg2}	0.26287	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{gg2}	31.702	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$
Grout to soil	R_{gs}	0.05861	$\mathrm{m\,s\,K\,J^{-1}}$	Φ_{gs}	181.02	$\mathrm{J\,m^{-2}\,s^{-1}\,K^{-1}}$

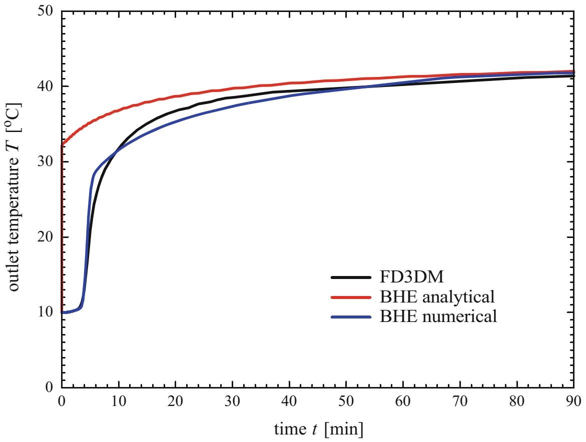


Fig. 13.16 Short-term outlet temperature history of the BHE solution in comparison to the FD3DM solution measured at the pipe's outlet

For the FD3DM the GLS predictor-corrector AB/TR time integration scheme with a RMS error tolerance of 10^{-4} has been used. It took 276 time steps for the simulation period of 365 days. For the BHE solutions always the FE/BE time

Fig. 13.17 Long-term outlet temperature history of the BHE solution in comparison to the FD3DM solution measured at the pipe's outlet

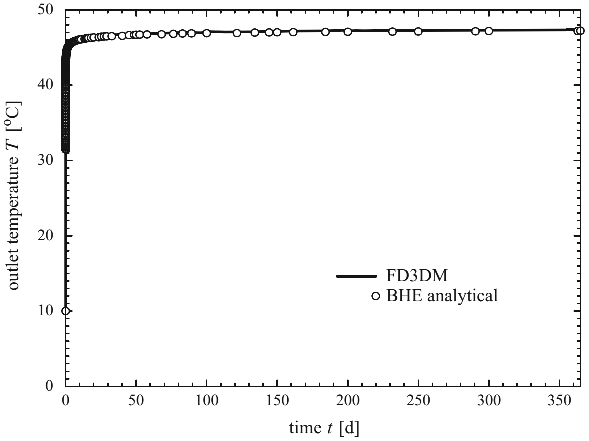
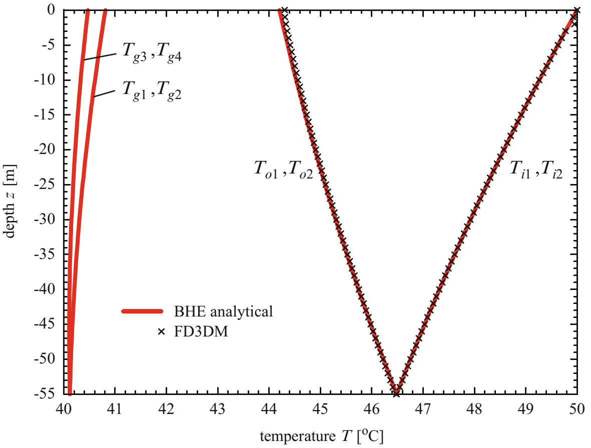


Fig. 13.18 Analytical BHE solution of temperature profile at $t = 12\,\mathrm{h}$ in comparison to the FD3DM solution



marching predictor-corrector scheme with a RMS error tolerance of 10^{-3} was preferred due to better robustness for this class of problems. The analytical BHE required only 227 time steps.

Chapter 14
Discrete Feature Modeling of Flow, Mass and Heat Transport Processes

14.1 Introduction

The discrete feature approach provides the crucial link between the complex geometries for subsurface and surface, porous and fractured continua as well as to incorporate engineered structures in modeling flow, mass and heat transport processes. In such a *holistic* approach a 3D geometry of the subsurface domain (aquifer system, rock masses) in describing a porous-medium structure can be combined by interconnected 1D and/or 2D discrete features as shown in Fig. 14.1. In the finite element context the 3D mesh for the porous medium can be enriched by discrete line (channel, borehole, pipe network, tunnel, mine stope) and/or areal (overland, fault, fracture) elements.

Discrete features are geometric representations of a lower spatial dimension having commonly a significant fluid conductance in comparison to the porous medium. Their conceptual modeling approach and resulting basic equations are thoroughly described in Chap. 4, where a unified basis in form of *diffusion-type* flow equations, e.g., Darcy, Hagen-Poiseuille or Manning-Strickler laws of fluid motion, as well as mass and heat transport equations is derived. Discrete features are approximated as 1D or 2D finite elements termed as *discrete feature elements* (DFE's), which can be mixed with porous-medium elements in two and three dimensions (Fig. 14.2). The 1D and 2D DFE's share the nodes of the porous-medium elements and can be placed along element edges and faces or can even interconnect arbitrary nodes in a finite element mesh. Both the geometric and physical characteristics of DFE provide a large flexibility in modeling complex situations.

14.2 Discrete Feature Master Equations

The governing balance equations for discrete features as listed in Tables 4.5–4.7 for flow, species mass and heat, respectively, can be generalized by the following advection-diffusion-type master equation

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_14,

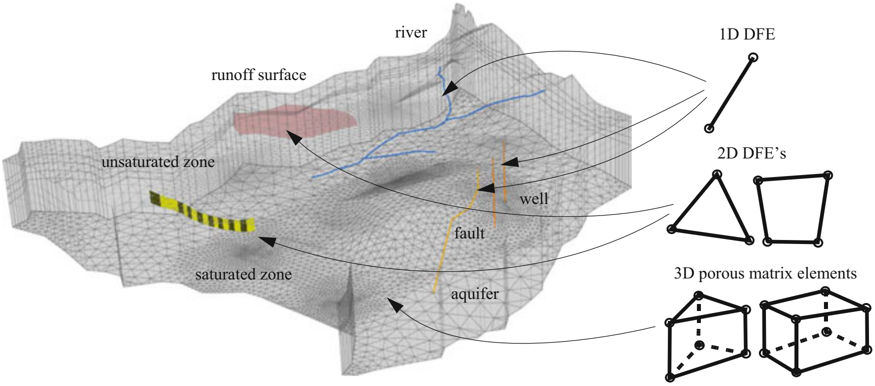
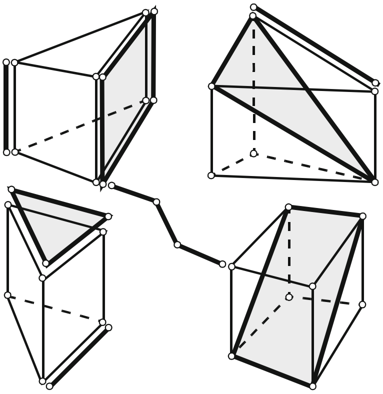


Fig. 14.1 Schematization of a subsurface modeling system by combining discrete features with volume discretizations of the total study domain: 1D feature elements are used to approximate rivers, channels, wells and specific faults, 2D feature elements are appropriate for modeling runoff processes, fractured surfaces and faulty zones, and 3D elements represent the basic tessellation of the subsurface domain consisting of an aquifer-aquitard system and involving unsaturated and saturated zones

Fig. 14.2 Example of 1D and 2D DFE's mixed with 3D porous-medium elements



$$\mathcal{L}_F(\phi) = \mathcal{S}_F \frac{\partial \phi}{\partial t} + \boldsymbol{v}_F \cdot \nabla\phi - \nabla \cdot (\boldsymbol{\Upsilon}_F \cdot \nabla\phi) + \Theta_F \phi - Q_F + Q_{\phi w_F} = 0 \qquad (14.1)$$
$$\text{in} \quad \Omega_F \subset \Re^D \ (D = 1, 2), \ t \geq t_0$$

which has to be solved for flow ($\phi := h$), species mass ($\phi := C_k$) and heat ($\phi := T$) in 1D or 2D space $\Omega_F \subset \Omega$ of a discrete feature F subject to the Dirichlet, Neumann and Cauchy BC's as well as well-type SPC as

$$\begin{array}{ll} \phi = \phi_D & \text{on } \Gamma_{D_F} \times t[t_0,\infty) \\ -(\boldsymbol{\Upsilon}_F \cdot \nabla\phi)\cdot \boldsymbol{n} = q_N & \text{on } \Gamma_{N_F} \times t[t_0,\infty) \\ -(\boldsymbol{\Upsilon}_F \cdot \nabla\phi)\cdot \boldsymbol{n} = -\Phi_F(\phi_C - \phi) & \text{on } \Gamma_{C_F} \times t[t_0,\infty) \\ Q_{\phi w_F} = -\sum_w \big(\phi_w - \phi(\boldsymbol{x}_w)\big) Q_w(t)\delta(\boldsymbol{x}-\boldsymbol{x}_w) & \text{on } \boldsymbol{x}_w \in \Omega_F \times t[t_0,\infty) \end{array} \quad (14.2)$$

imposed on the discrete feature boundary $\Gamma_F = \Gamma_{D_F} \cup \Gamma_{N_F} \cup \Gamma_{C_F}$ and associated with IC of the form

$$\phi(\boldsymbol{x},t_0) = \phi_0(\boldsymbol{x}) \quad \text{in} \quad \bar{\Omega}_F \quad (14.3)$$

In the master equation (14.1) ϕ corresponds to a generalized variable, the gradient operator ∇ refers only to 1D or 2D space and $\mathcal{S}_F$, $\boldsymbol{v}_F$, $\boldsymbol{\Upsilon}_F$, Θ_F, Q_F, $Q_{\phi w_F}$ represent specific quantities of storage, advection, dispersion, transfer, supply and well-type sink/source, respectively, which have to be specified from Tables 4.5–4.7 in dependence on the occurring type and dimension of the discrete feature F. The essential parameters required for solving (14.1) with (14.2) and (14.3) are listed in Table I.9 for flow, in Table I.15 for mass transport and in Table I.18 for heat transport of Appendix I. *Steady-state* situations occur if $\mathcal{S}_F = 0$.[1]

14.3 Finite Element Formulation

The governing ADE (14.1) of flow, mass and heat transport in 1D or 2D discrete features is mathematically similar to the paradigmatic ADE of a scalar variable used in Chap. 8. Based on the principles given there we use now the GFEM to solve (14.1) for the generalized variable $\phi = \phi(\boldsymbol{x},t)$ subject to the corresponding BC's (14.2) and IC (14.3). Since most of the details are equivalent to the ADE developments given in Chap. 8 we shall focus here only on the specific aspects related to the DFE approach.

14.3.1 Weak Form

In analogy to the statement (8.55) of Sect. 8.5 we find the corresponding weak form for the governing ADE (14.1) in its convective form as

[1] Optionally, FEFLOW suppresses the time derivative term $\partial\phi/\partial t$ for solving steady-state solutions.

$$\int_{\Omega_F} w\mathcal{S}_F \frac{\partial \phi}{\partial t} d\Omega + \int_{\Omega_F} w\boldsymbol{v}_F \cdot \nabla \phi d\Omega + \int_{\Omega_F} \nabla w \cdot (\boldsymbol{\Upsilon}_F \cdot \nabla \phi) d\Omega + \int_{\Omega_F} w[\Theta_F \phi - Q_F] d\Omega + \sum_w w(\boldsymbol{x}_w)\big(\phi_w - \phi(\boldsymbol{x}_w)\big) Q_w(t) + \int_{\Gamma_{N_F}} wq_N d\Gamma - \int_{\Gamma_{C_F}} w\Phi_F(\phi_C - \phi) d\Gamma = 0, \quad \forall w \in H_0^1(\Omega_F) \tag{14.4}$$

where w is a suitable weighting function and the boundary integrals are suitably separated into their segments $\Gamma_F = \Gamma_{D_F} \cup \Gamma_{N_F} \cup \Gamma_{C_F}$ imposed by the Dirichlet, Neumann and Cauchy-type BC's (14.2) on the discrete feature.

14.3.2 GFEM and Resulting Matrix System

In using the FEM the unknown variable ϕ appearing in the weak statement (14.4) is replaced by a *continuous approximation* that assumes the separability of space and time (cf. Sect. 8.4). Thus

$$\phi(\boldsymbol{x}, t) \approx \sum_j N_j(\boldsymbol{x})\phi_j(t), \quad j = 1, \ldots, N_{\mathrm{P}_F} \tag{14.5}$$

where j designates global nodal indices and N_{P_F} is the number of nodal points related to a discrete feature F. Using the Galerkin method with the weighting function

$$w \rightarrow w_i = N_i, \quad i = 1, \ldots, N_{\mathrm{P}_F} \tag{14.6}$$

and applying the approximate solutions (14.5) in (14.4), the following matrix system of N_{P_F} equations results

$$\boldsymbol{O}_F \cdot \dot{\boldsymbol{\phi}} + \boldsymbol{K}_F \cdot \boldsymbol{\phi} - \boldsymbol{F}_F = \boldsymbol{0} \tag{14.7}$$

where

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N_{\mathrm{P}_F}} \end{pmatrix}, \quad \dot{\boldsymbol{\phi}} = \begin{pmatrix} \frac{d\phi_1}{dt} \\ \frac{d\phi_2}{dt} \\ \vdots \\ \frac{d\phi_{N_{\mathrm{P}_F}}}{dt} \end{pmatrix} \tag{14.8}$$

and the matrices and RHS vector

$$\begin{aligned}
\boldsymbol{O}_F &= O_{ij_F} = \sum_e \int_{\Omega_F^e} \mathcal{S}_F^e\, N_i N_j \, d\Omega^e \\
\boldsymbol{K}_F &= K_{ij_F} = \sum_e \Big(\int_{\Omega_F^e} N_i \boldsymbol{v}_F^e \cdot \nabla N_j \, d\Omega^e + \int_{\Omega_F^e} \nabla N_i \cdot (\boldsymbol{\Upsilon}_F^e \cdot \nabla N_j) d\Omega^e + \\
&\qquad \int_{\Omega_F^e} \Theta_F^e N_i N_j \, d\Omega^e + \int_{\Gamma_{C_F}^e} \Phi_F^e N_i N_j \, d\Gamma^e \Big) - \delta_{ij} Q_w(t) \Big|_i \\
\boldsymbol{F}_F &= F_{i_F} = \sum_e \Big(\int_{\Omega_F^e} N_i Q_F^e d\Omega^e + \int_{\Gamma_{C_F}^e} N_i \Phi_F^e \phi_C^e \, d\Gamma^e - \int_{\Gamma_{N_F}^e} N_i q_N^e \, d\Gamma^e \Big) - \phi_w Q_w(t) \Big|_i
\end{aligned} \tag{14.9}$$

where $(i, j = 1, \ldots, N_{\mathrm{P}_F})$ and $(e = 1, \ldots, N_{\mathrm{E}_F})$. The integrals appearing in (14.9) are integrated on element level in the local coordinates $\boldsymbol{\eta}$ as described in Sect. 8.12. Analytical evaluations of partial integral terms of (14.9) can be deduced from developments done in Appendix H for selected element types, in particular for the 1D linear line element (Sect. H.1) and the 2D linear triangular element (Sect. H.2). Numerical integration via Gauss-Legendre quadrature is employed for 2D quadrilateral DFE's.

14.3.3 Assembly of DFE's into the Global System Matrix

14.3.3.1 Need for Coordinate Transformation

The matrix system (14.7) written in the form

$$\begin{aligned}
\boldsymbol{O}_F \cdot \dot{\boldsymbol{\phi}} + \boldsymbol{K}_F \cdot \boldsymbol{\phi} &= \boldsymbol{F}_F \\
\boldsymbol{O}_F &= \textstyle\sum_e \boldsymbol{O}_F^e \\
\boldsymbol{K}_F &= \textstyle\sum_e \boldsymbol{K}_F^e \\
\boldsymbol{F}_F &= \textstyle\sum_e \boldsymbol{F}_F^e
\end{aligned} \tag{14.10}$$

provides elemental matrix and vector contributions for each DFE e of feature F, which must be assembled additionally into the global finite-element matrix systems resulting from the porous-media equations developed in the preceding sections, such as (9.20) for flow in saturated porous media, (10.30) for flow in variably saturated porous media, (11.38) for variable-density flow in porous media, (12.22) for mass transport in porous media and (13.22) for heat transport in porous media.

The integrals (14.9) in $\boldsymbol{O}_F^e$, $\boldsymbol{K}_F^e$ and $\boldsymbol{F}_F^e$ for each DFE e and feature F are performed in the local coordinates $\boldsymbol{\eta}$ for the corresponding Euclidean space $\Re^D$ (cf. Sects. 8.8 and 8.11). Usually, 1D finite elements are mapped to the $\Re^1$ space, 2D elements to the $\Re^2$ space and 3D elements to the $\Re^3$ space. In such cases the mapping is strictly one-to-one, that means three global coordinates (x, y, z) are transformed to three local coordinates (ξ, η, ζ) in 3D, two global coordinates (x, y) to two local coordinates (ξ, η) in 2D and one global coordinate (x) to one local coordinate (ξ) in 1D. However, when 1D and 2D DFE's are generally mapped onto a 3D global space, the number of local coordinates $\boldsymbol{\eta}$ will be less than the number of global coordinates $\boldsymbol{x}^e$ and the transformation Jacobian $\boldsymbol{J}^e = \partial \boldsymbol{x}^e / \partial \boldsymbol{\eta}$ for finite elements (8.115) will not be any more an invertible square matrix (e.g., for the $\xi - \eta-$system of a 2D DFE mapped onto the global $x - y - z-$system the third row of $\boldsymbol{J}^e$ contains zeros, $J_{31}^e = J_{32}^e = J_{33}^e = 0$, because the $\zeta-$coordinate does not exist in 2D elements).

There are at least two ways to overcome this mapping conflict. A more general method has been proposed by Perrochet [414], who uses expressions of gradients in curvilinear coordinates and introduces covariant bases and metric tensors to replace the usual Jacobian. Alternatively, the method of coordinate transformation appears as a cost-effective and simpler method, we shall prefer here. Taking into consideration that all flow and transport processes are invariant with respect to a rotation (orthogonal transformation) of the global coordinates $\boldsymbol{x} = \boldsymbol{x}^e$, $\forall e$, we can arbitrarily rotate $\boldsymbol{x}$ to the $\boldsymbol{x}'^e-$coordinates different for each element e by using a suitable rotation matrix of directional cosines $\boldsymbol{A}^e$ as

$$\begin{aligned} \boldsymbol{x}'^e &= \boldsymbol{A}^e \cdot \boldsymbol{x}^e \\ \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}^e &= \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}^e \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix}^e \end{aligned} \tag{14.11}$$

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 DFE's in the $\Re^3$ 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. 14.3).

The advantages of this coordinate transformation are that the corresponding Jacobian $\boldsymbol{J}'^e$

$$\boldsymbol{J}'^e = \frac{\partial \boldsymbol{x}'^e}{\partial \boldsymbol{\eta}} \tag{14.12}$$

becomes again an invertible square matrix and the standard metric procedure can be maintained in the assembly process for the global matrix system (14.10). To ease the computations the $x' - y' - z'-$coordinate system may, in fact, be different for every element e. Actually, the integrals (14.9) in $\boldsymbol{O}^e$, $\boldsymbol{K}^e$ and $\boldsymbol{F}^e$ over Ω_F^e, $\Gamma_{C_F}^e$ and $\Gamma_{N_F}^e$ are performed in the local coordinates $\boldsymbol{\eta}$ which are directly mapped onto the transformed coordinates $\boldsymbol{x}'^e$ (Fig. 14.3):

$$\Omega_F^e = \Omega_F^e(\boldsymbol{x}'^e(\boldsymbol{\eta})), \quad \Gamma_{C_F}^e = \Gamma_{C_F}^e(\boldsymbol{x}'^e(\boldsymbol{\eta})), \quad \Gamma_{N_F}^e = \Gamma_{N_F}^e(\boldsymbol{x}'^e(\boldsymbol{\eta})) \tag{14.13}$$

14.3.3.2 Determination of the Directional Cosines $\boldsymbol{A}^e$ of DFE e

The directional cosines $\boldsymbol{A}^e$ are only required for mapping 2D and 1D DFE's in the $\Re^3$ space. Suppose the 3D continuum domain Ω with its boundary Γ is completely filled by 3D finite elements (e.g., hexahedral or pentahedral isoparametric elements),

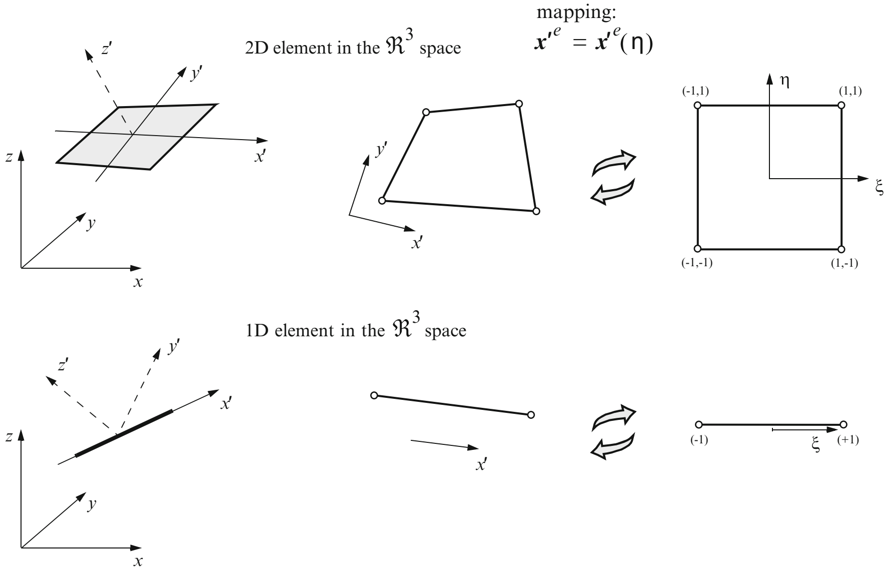


Fig. 14.3 Global $x - y - z$—coordinate system, rotated elemental $x' - y' - z'$—coordinate system and local ξ — (η)—coordinate system for 2D and 1D DFE's in the $\Re^3$ space

the 1D and 2D DFE's share the nodal points of the 3D mesh and their geometric extents are aligned to surfaces, edges or diagonals of the 3D porous-medium elements (Fig. 14.4).

For 2D DFE's forming surfaces of the 3D porous-medium element it is convenient to derive the directional cosines directly from the shape of the 3D element. We can construct the two directional vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ (Fig. 14.4), which are parallel to the local ξ— and η—axes, respectively. They can be found by the following shape-derived relationships

$$\boldsymbol{u}_1 = \begin{cases} \begin{pmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \\ \frac{\partial z}{\partial \xi} \end{pmatrix} = \begin{pmatrix} J_{11} \\ J_{12} \\ J_{13} \end{pmatrix}^e & \text{at} \quad \zeta = \pm 1 \\ \begin{pmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \\ \frac{\partial z}{\partial \eta} \end{pmatrix} = \begin{pmatrix} J_{21} \\ J_{22} \\ J_{23} \end{pmatrix}^e & \text{at} \quad \xi = \pm 1 \\ \begin{pmatrix} \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \zeta} \end{pmatrix} = \begin{pmatrix} J_{31} \\ J_{32} \\ J_{33} \end{pmatrix}^e & \text{at} \quad \eta = \pm 1 \end{cases} \tag{14.14}$$

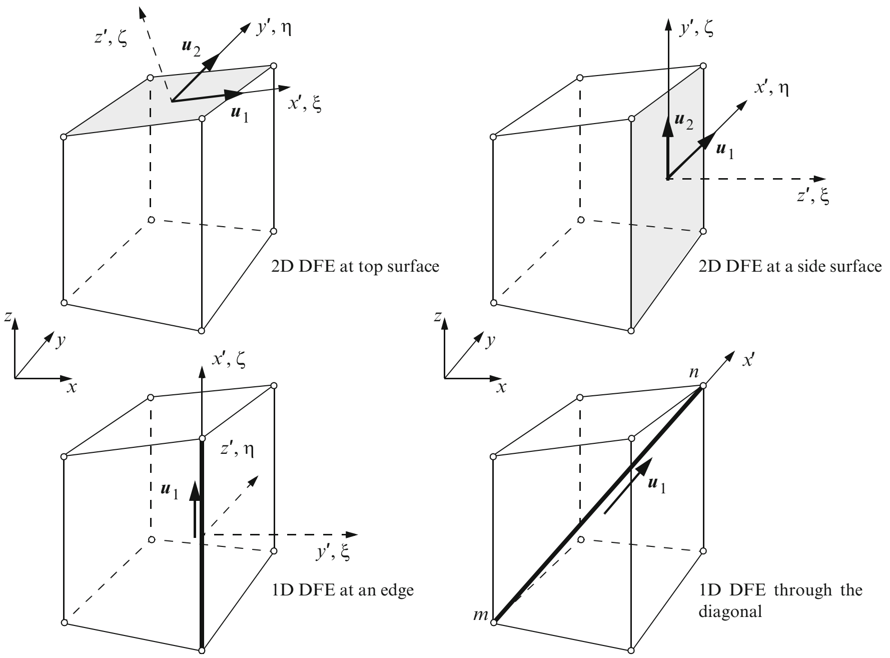


Fig. 14.4 Exemplified mapping of 2D and 1D DFE's aligned to surfaces, edges and diagonals, respectively, for a 3D finite porous-medium element. Global and local coordinates

$$
\boldsymbol{u}_2 = \begin{cases} \begin{pmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \\ \frac{\partial z}{\partial \eta} \end{pmatrix} = \begin{pmatrix} J_{21} \\ J_{22} \\ J_{23} \end{pmatrix}^e & \text{at} \quad \zeta = \pm 1 \\ \begin{pmatrix} \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \zeta} \end{pmatrix} = \begin{pmatrix} J_{31} \\ J_{32} \\ J_{33} \end{pmatrix}^e & \text{at} \quad \xi = \pm 1 \\ \begin{pmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \\ \frac{\partial z}{\partial \xi} \end{pmatrix} = \begin{pmatrix} J_{11} \\ J_{12} \\ J_{13} \end{pmatrix}^e & \text{at} \quad \eta = \pm 1 \end{cases} \tag{14.15}
$$

These directional vectors can be easily used to compute the directional cosines according to

$$
A_{ij}^e = \cos(\boldsymbol{u}_i, \boldsymbol{e}_j) = \frac{\boldsymbol{u}_i \cdot \boldsymbol{e}_j}{\|\boldsymbol{u}_i\| \underbrace{\|\boldsymbol{e}_j\|}_{=1}} \quad \text{for} \quad \begin{aligned} i &= 1,2 \\ j &= 1,2,3 \end{aligned} \tag{14.16}
$$

with the base vectors (2.5):

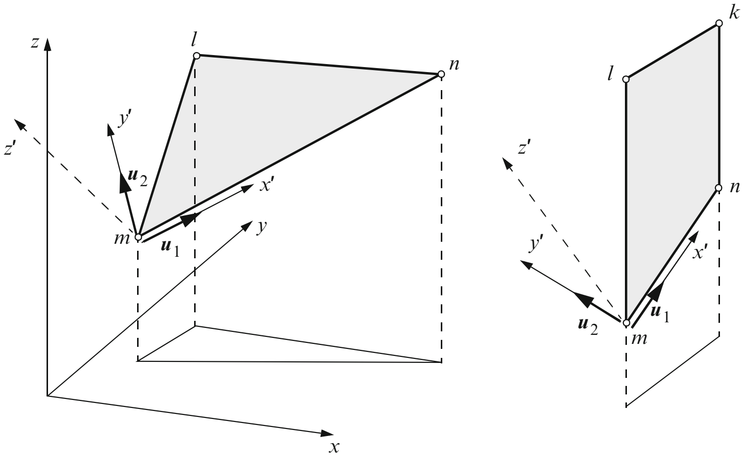


Fig. 14.5 Directional vectors $\boldsymbol{u}_i$ $(i = 1, 2)$ for a linear triangular element and a vertically-oriented linear quadrilateral element

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{14.17}$$

Note that for 2D DFE's we need only two directional vectors $(i = 1, 2)$, the remaining directional cosines A^e_{3j} are meaningless.

Often we can assume that the 2D DFE's are perfectly *plane*, 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 (14.14) and (14.15), in such cases it is convenient to derive the directional vectors $\boldsymbol{u}_i$ in a direct manner as follows (see Fig. 14.5). We specify the $x'-$axis along the edge nm of the 2D DFE. The vector $\boldsymbol{u}_1$ is accordingly given by

$$\boldsymbol{u}_1 = \begin{pmatrix} x_n - x_m \\ y_n - y_m \\ z_n - z_m \end{pmatrix} \tag{14.18}$$

The second directional vector $\boldsymbol{u}_2$ derived by simple vector algebra yields[2]

$$\boldsymbol{u}_2 = \boldsymbol{q} - \Big(\frac{\boldsymbol{q} \cdot \boldsymbol{u}_1}{\boldsymbol{u}_1 \cdot \boldsymbol{u}_1}\Big)\boldsymbol{u}_1 \tag{14.19}$$

with the auxiliary vector $\boldsymbol{q}$ formed along the adjacent side lm of the 2D element as

[2] Since the vector $\boldsymbol{u}_2 = \boldsymbol{q} - \boldsymbol{v}$ is perpendicular to the vector $\boldsymbol{u}_1$, the dot product yields

$$\boldsymbol{q} = \begin{pmatrix} x_l - x_m \\ y_l - y_m \\ z_l - z_m \end{pmatrix} \tag{14.20}$$

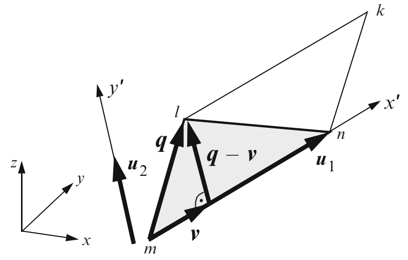
and the directional cosines A^e_{ij} $(i = 1, 2; j = 1, 2, 3)$ can be easily computed by using (14.16).

For 1D DFE's the same procedure can be applied to determine A^e_{ij} for $(i = 1; j = 1, 2, 3)$. Here, only one row A^e_{1j} of the rotation matrix is of interest. Taking into consideration that 1D DFE's 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^e_{1j} for a 1D linear (noncurved) DFE spanning between the two nodes n and m (cf. Fig. 14.4):

$$\boldsymbol{u}_1 = \begin{pmatrix} x_n - x_m \\ y_n - y_m \\ z_n - z_m \end{pmatrix} = \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix}, \qquad \|\boldsymbol{u}_1\| = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \tag{14.21}$$

$$\begin{aligned} A^e_{11} &= \frac{\Delta x}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \\ A^e_{12} &= \frac{\Delta y}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \\ A^e_{13} &= \frac{\Delta z}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \end{aligned} \tag{14.22}$$

$$(\boldsymbol{q} - \boldsymbol{v}) \cdot \boldsymbol{u}_1 = 0$$



Using a parametric description of the vector $\boldsymbol{v} = t\boldsymbol{u}_1$, the parameter t can be easily found:

$$\begin{aligned} (\boldsymbol{q} - t\boldsymbol{u}_1) \cdot \boldsymbol{u}_1 &= 0 \\ t &= \tfrac{\boldsymbol{q}\cdot\boldsymbol{u}_1}{\boldsymbol{u}_1\cdot\boldsymbol{u}_1} \end{aligned}$$

As a result, we get $\boldsymbol{v} = \big(\tfrac{\boldsymbol{q}\cdot\boldsymbol{u}_1}{\boldsymbol{u}_1\cdot\boldsymbol{u}_1}\big)\boldsymbol{u}_1$ and finally $\boldsymbol{u}_2 = \boldsymbol{q} - \big(\tfrac{\boldsymbol{q}\cdot\boldsymbol{u}_1}{\boldsymbol{u}_1\cdot\boldsymbol{u}_1}\big)\boldsymbol{u}_1$.

14.3.4 Implicit Monolithic Solution of the Coupled Discrete Fracture and Porous-Medium Equations

Discrete features and the porous medium are treated as a monolithic entity, where all components are implicitly integrated in the solution domain

$$\Omega = \Omega_P \cup \sum_F \Omega_F \tag{14.23}$$

consisting of the joint porous-medium domain Ω_P and a number of nonoverlapping discrete feature domains Ω_F, governed by different balance equations, however, solvable via a common state variable $\phi = \phi(\boldsymbol{x}, t)$ (e.g., h for flow, C_k for species mass and T for heat transport). In the finite element context the elementwise continuous approximation for $\phi \approx \hat{\phi}$ allows the assembly of the elemental contributions of porous medium and discrete features in a standard manner such that (cf. Sect. 8.6)

$$\begin{aligned} \textstyle\int_\Omega \{\ldots\} d\Omega &= \sum_e \int_{\Omega^e} \{\ldots\} d\Omega^e \\ &= \sum_e \Big(\int_{\Omega_P^e} w\mathcal{L}_P(\hat{\phi}) d\Omega^e + \sum_F \int_{\Omega_F^e} w\mathcal{L}_F(\hat{\phi}) d\Omega^e \Big) \end{aligned} \tag{14.24}$$

where $\mathcal{L}_P(\hat{\phi})$ and $\mathcal{L}_F(\hat{\phi})$ represent the governing PDE's for the porous medium and the discrete features, respectively. Practically, the formation of the global matrix system comprising both the porous medium and the discrete feature entities is simply the assembly of their partial matrix and vector contributions. For example, the matrix system of the porous-medium flow (9.20) extends now to

$$\boldsymbol{M} \cdot \dot{\boldsymbol{h}} + \boldsymbol{D} \cdot \boldsymbol{h} - \boldsymbol{Z} = \boldsymbol{0} \tag{14.25}$$

with

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{O} + \textstyle\sum_F \boldsymbol{O}_F \\ \boldsymbol{D} &= \boldsymbol{C} + \textstyle\sum_F \boldsymbol{K}_F \\ \boldsymbol{Z} &= \boldsymbol{F} + \textstyle\sum_F \boldsymbol{F}_F \end{aligned} \tag{14.26}$$

to solve the hydraulic head $\boldsymbol{h}$, where $\boldsymbol{O}$, $\boldsymbol{C}$ and $\boldsymbol{F}$ are the porous-medium contributions given by (9.22) and $\boldsymbol{O}_F$, $\boldsymbol{K}_F$ and $\boldsymbol{F}_F$ are the discrete feature F contributions given by (14.9). Similarly, for variably saturated flow, variable-density flow, species mass and heat transport the global matrix systems result if we assemble the corresponding porous-medium matrix system (10.30), (11.38), (12.22) and (13.22), respectively, with the matrix system (14.7) for the discrete feature F and associated state variable $\boldsymbol{h}$, $\boldsymbol{C}_k$ and $\boldsymbol{T}$ of hydraulic head, species concentration and temperature, respectively.

Since the DFE's share the same nodal points with the porous medium, a natural result of the assembly process is in the *parallel* behavior of exchanging (advective and conductive/diffusive) fluxes between porous medium and discrete features. Suppose K_P and K_F represent characteristic conductivities of porous medium and discrete feature, respectively, at the same node, an exchanging flux between porous medium and discrete feature is affected by its effective conductivity $K = K_P + K_F$. If $K_F \gg K_P$ the flux becomes dominated by the discrete feature property, however if $K_F \rightarrow 0$ the effect from DFE vanishes and the exchanging flux is determined by the porous-medium property alone. A disadvantageous consequence of the latter is that there is no possibility to model clogging or sealing effects by simply using DFE's with sharing nodes because the exchanging flux can never be smaller than that of the porous medium since $K \geq K_P$, except for $K_P \rightarrow 0$.

14.3.5 Time Integration

The global matrix systems for flow such as (14.25), and similar to mass and heat transport comprising both the porous medium and the discrete feature entities have to be solved in time t with the associated IC's via suitable single-step semi-implicit or fully implicit time marching recurrence schemes as described in Sect. 8.13. The GLS predictor-corrector time stepping method combined with an automatic error-controlled time step selection strategy is usually preferred. Its solution steps applied to the global matrix systems are fully equivalent to the procedures as thoroughly described above in Sect. 8.13.5 (summarized in Table 8.7) for a general ADE, in Sect. 10.7.5 for unsaturated flow, in Sect. 11.6.4 for density-variable flow, in Sect. 12.3.3 for reactive mass transport and in Sect. 13.3.3 for heat transport.

14.4 Computation of Velocity Fields and Budget Analysis

The flow vectors for the porous medium $\boldsymbol{q}_P$ and for the discrete features $\boldsymbol{v}_F$ are at first separately evaluated by using smoothing techniques as thoroughly described in Sect. 8.19.1. For the porous medium continuous Darcy velocities $\boldsymbol{q}_P$ at the nodal points are derived such as given in (10.120) for variably saturated media and in (11.69) for variable-density flow. The same procedures are applied to compute the discrete velocities at the nodes of a discrete feature F. For example, for the Hagen-Poiseuille flow velocity (4.51) and for the overland and channel flow velocity (4.63) the following discrete evaluation is performed

$$\boldsymbol{v}_F(\boldsymbol{x}, t_{n+1}) = -\sum_j \boldsymbol{K}_F f_\mu \cdot \left[\nabla N_j(\boldsymbol{x})\, h_j(t_{n+1}) + \chi \boldsymbol{e}\right] \tag{14.27}$$

by using the known hydraulic head values $h_j(t_{n+1}) = \boldsymbol{h}_{n+1}$ at nodes j of discrete feature F at time plane $n+1$, where $\boldsymbol{K}_F$ corresponds to a generalized hydraulic conductivity of discrete feature F specifying the different flow laws according to (4.51) or (4.63). Note that the velocity $\boldsymbol{v}_F$ is only smoothed separately for the contributions of the discrete feature nodes j, however, no smoothing is performed with the velocity contributions of the porous medium. Finally, the total velocity $\boldsymbol{q}$ is a result of superimposing the velocity of porous medium and discrete feature at given location $\boldsymbol{x}$ and time stage t_{n+1}, viz.,

$$\boldsymbol{q}(\boldsymbol{x}, t_{n+1}) = \boldsymbol{q}_P(\boldsymbol{x}, t_{n+1}) + \boldsymbol{v}_F(\boldsymbol{x}, t_{n+1}) \tag{14.28}$$

Note, however, the evaluations of the advective terms in the corresponding mass and heat transport equations are always based on their separate (nonsuperimposed) flow fields, i.e., porous-medium equations take the porous-medium Darcy velocities $\boldsymbol{q}_P$ and fracture equations use the fracture velocities $\boldsymbol{v}_F$.

The precise budget analysis for flow, mass and heat transport problems which are mixed with discrete features is fully analogous to the technique as described for porous-media processes in Sects. 9.7, 10.11, 11.8, 12.4 and 13.4 based on CBFM introduced and thoroughly described in Sect. 8.19.2. Similarly, we use the basic weak statement (14.4) of the discrete feature transport equation to express the corresponding boundary flux on the discrete feature boundary Γ_F as

$$\begin{aligned}\int_{\Gamma_F} N_i\, q_{n_F}\, d\Gamma = -\int_{\Omega_F} N_i \mathcal{S}_F \frac{\partial \phi}{\partial t} d\Omega - \int_{\Omega_F} N_i \boldsymbol{v}_F \cdot \nabla \phi d\Omega - \\ \int_{\Omega_F} \nabla N_i \cdot (\boldsymbol{\Upsilon}_F \cdot \nabla \phi) d\Omega - \int_{\Omega_F} N_i (\Theta_F \phi - Q_F) d\Omega - (\phi_w - \phi) Q_w(t)|_i\end{aligned} \tag{14.29}$$

which leads to the matrix system

$$\boldsymbol{M}_F \cdot \boldsymbol{q}_{n_F} = -\boldsymbol{O}_F \cdot \dot{\boldsymbol{\phi}} - \boldsymbol{K}_F^{\dagger} \cdot \boldsymbol{\phi} + \boldsymbol{F}_F^{\dagger} \tag{14.30}$$

with

$$\begin{array}{rcll}\boldsymbol{M}_F &=& M_{ij_F} &= \int_{\Gamma} N_i N_j d\Gamma \\ \boldsymbol{O}_F &=& O_{ij_F} &= \int_{\Omega_F} \mathcal{S}_F N_i N_j d\Omega \\ \boldsymbol{K}_F^{\dagger} &=& K_{ij_F}^{\dagger} &= \int_{\Omega_F} N_i \boldsymbol{v}_F \cdot \nabla N_j d\Omega + \int_{\Omega_F} \nabla N_i \cdot (\boldsymbol{\Upsilon}_F \cdot \nabla N_j) d\Omega + \\ &&& \quad \int_{\Omega_F} \Theta_F N_i N_j d\Omega - \delta_{ij} Q_w(t)\Big|_i \\ \boldsymbol{F}_F^{\dagger} &=& F_{i_F}^{\dagger} &= \int_{\Omega_F} N_i Q_F d\Omega - \phi_w Q_w(t)\Big|_i \end{array} \tag{14.31}$$

for solving the continuous boundary flux vector $\boldsymbol{q}_{n_F}$ of discrete feature F. To compute the boundary flux of the total system comprising both the porous medium and the discrete feature entities we can simply assembly their partial matrix and vector contributions. For example, the matrix system of porous-medium flow budget (9.65) extends now to

$$\boldsymbol{M}^{\ddagger} \cdot \boldsymbol{q}_n^{\ddagger} = -\boldsymbol{O}^{\ddagger} \cdot \dot{\boldsymbol{h}} - \boldsymbol{C}^{\ddagger} \cdot \boldsymbol{h} + \boldsymbol{F}^{\ddagger} \tag{14.32}$$

with

$$\begin{aligned} \boldsymbol{q}_n^{\ddagger} &= \boldsymbol{q}_n + \textstyle\sum_F \boldsymbol{q}_{n\,F} \\ \boldsymbol{M}^{\ddagger} &= \boldsymbol{M} + \textstyle\sum_F \boldsymbol{M}_F \\ \boldsymbol{O}^{\ddagger} &= \boldsymbol{O}^{\dagger} + \textstyle\sum_F \boldsymbol{O}_F \\ \boldsymbol{C}^{\ddagger} &= \boldsymbol{C}^{\dagger} + \textstyle\sum_F \boldsymbol{K}_F^{\dagger} \\ \boldsymbol{F}^{\ddagger} &= \boldsymbol{F}^{\dagger} + \textstyle\sum_F \boldsymbol{F}_F^{\dagger} \end{aligned} \tag{14.33}$$

to solve the total boundary flux vector $\boldsymbol{q}_n^{\ddagger}$, where $\boldsymbol{M}$, $\boldsymbol{O}^{\dagger}$, $\boldsymbol{C}^{\dagger}$ and $\boldsymbol{F}^{\dagger}$ are the porous-medium contributions given by (9.66) and $\boldsymbol{M}_F$, $\boldsymbol{O}_F$, $\boldsymbol{K}_F^{\dagger}$ and $\boldsymbol{F}_F^{\dagger}$ are the discrete feature F contributions given by (14.31). Note that in the budget analysis the total integral flux $\boldsymbol{Q}_n^{\ddagger}$ is directly evaluated at each boundary node by

$$\begin{aligned} \boldsymbol{Q}_n^{\ddagger} &= -\boldsymbol{M}^{\ddagger} \cdot \boldsymbol{q}_n^{\ddagger} \\ &= \boldsymbol{O}^{\ddagger} \cdot \dot{\boldsymbol{h}} + \boldsymbol{C}^{\ddagger} \cdot \boldsymbol{h} - \boldsymbol{F}^{\ddagger} \end{aligned} \tag{14.34}$$

Similarly, for variably saturated flow, variable-density flow, species mass and heat transport the global matrix systems for budget analysis result if assembly the corresponding porous-medium matrix system (10.123), (11.73), (12.47) and (13.30), respectively, with the matrix system (14.30) for the discrete feature F and associated state variable $\boldsymbol{h}$, $\boldsymbol{C}_k$ and $\boldsymbol{T}$ of hydraulic head, species concentration and temperature, respectively.

14.5 Examples

14.5.1 Solute Diffusion into Porous Matrix from a Single Fracture

The single solute transport through fractured media was studied by Grisak and Pickens [215] for the case of a thin single fracture situated in a saturated porous rock as illustrated in Fig. 14.6. Advective transport is dominant in the fracture, while diffusive solute transport is usually dominant in the adjacent porous matrix. The diffusion into the porous matrix reduces the solute advancement in the fracture and thereby delays the migration of the solute, which acts as a diffusive loss for the fracture. Grisak and Pickens [215] used the standard FEM to model the fracture-matrix system, where the single fracture is discretized by thin areal 2D elements (i.e., no DFE).

An analytical solution for the fracture-matrix system of Fig. 14.6 has been developed by Tang et al. [506] by using Laplace transforms, which includes (1) advective transport along the fracture, (2) longitudinal dispersivity in the fracture, (3) molecular diffusion within the fracture, in the direction of the fracture

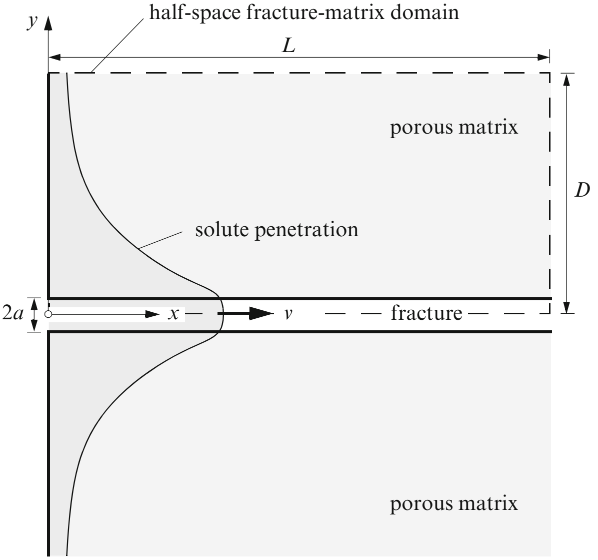


Fig. 14.6 Schematic sketch of the fracture-matrix system

axis x, (4) molecular diffusion from the fracture into the matrix, in the $y-$direction perpendicular to the fracture axis, (5) linear adsorption onto the face of the matrix, (6) linear adsorption within the matrix and (7) linear radioactive decay. It solves the coupled system of single solute mass balance equations governing in the fracture domain $(0 \leq x \leq \infty, 0 \leq y \leq a)$ as

$$\Re \frac{\partial C}{\partial t} + v \frac{\partial C}{\partial x} - D_{xx} \frac{\partial^2 C}{\partial x^2} + \Re \vartheta C - \left. \frac{\varepsilon D'_{yy}}{a} \frac{\partial C'}{\partial y} \right|_{y=a} = 0 \tag{14.35}$$

associated with the IC and BC's

$$C(x,0) = 0, \quad C(0,t) = C_D, \quad C(\infty,t) = 0 \tag{14.36}$$

and governing in the porous matrix domain $(a \leq y \leq \infty)$ as

$$\Re' \frac{\partial C'}{\partial t} - D'_{yy} \frac{\partial^2 C'}{\partial y^2} + \Re' \vartheta C' = 0 \tag{14.37}$$

associated with the IC and BC's

$$C'(x,y,0) = 0, \quad C'(x,a,t) = C(x,t), \quad C'(x,\infty,t) = 0 \tag{14.38}$$

with the retardation factors $\Re = 1 + \frac{K^{d'}}{a}$ and $\Re' = 1 + \frac{\rho_s K^d}{\varepsilon}$ as well as the dispersion coefficient $D_{xx} = D + \beta_L v = \mathcal{D}$ in the fracture and diffusion coefficient $D'_{yy} = \mathcal{D}'$ in the porous matrix, where all symbols quoted with $'$ refer to the porous matrix, unquoted symbols are related to the fracture or being indifferent, C and C' are the single solute concentrations in the fracture and in the porous matrix, respectively, a is the half of fracture width (see Fig. 14.6), $K^{d'}$ and K^{d} are the distribution coefficients for the porous matrix and fracture, respectively (cf. Table 3.8), ρ_s is the bulk density of the porous matrix, v is the groundwater velocity in the fracture (positive in $x-$direction), ε is the porosity of the porous matrix, ϑ is the decay rate and β_L is the longitudinal dispersivity.

Tang et al.'s general solution [506] of (14.35)–(14.38) takes the form of an integral which must be evaluated by numerical quadrature for each point in space and time (actually, Gaussian quadrature is used). On the other hand, a closed analytical transient solution can be derived for the simpler case which assumes negligible dispersion within the fracture, i.e., $\mathcal{D} \equiv 0$. It yields [506] the solute distribution within the fracture as

$$\begin{aligned}\frac{C}{C_D} = \frac{1}{2}\exp\Big(-\frac{\vartheta\Re x}{v}\Big)\Big[&\exp\Big(-\frac{\varepsilon\sqrt{\vartheta\Re'\mathcal{D}'}}{av}x\Big)\operatorname{erfc}\Big(\frac{\varepsilon\sqrt{\Re'\mathcal{D}'}}{2av\Re\sqrt{t-x\Re/v}}x-\sqrt{\vartheta}\sqrt{t-x\Re/v}\Big)+\\ &\exp\Big(\frac{\varepsilon\sqrt{\vartheta\Re'\mathcal{D}'}}{av}x\Big)\operatorname{erfc}\Big(\frac{\varepsilon\sqrt{\Re'\mathcal{D}'}}{2av\Re\sqrt{t-x\Re/v}}x+\sqrt{\vartheta}\sqrt{t-x\Re/v}\Big)\Big]\\ &\text{if}\ \ (t-x\Re/v)>0\\ &\text{otherwise}\ \ \frac{C}{C_D}=0\quad\text{if}\quad(t-x\Re/v)\le 0\end{aligned}\tag{14.39}$$

and the solute distribution within the porous matrix as

$$\begin{aligned}\frac{C'}{C_D} &= \frac{1}{2}\exp\Big(-\frac{\vartheta\Re x}{v}\Big)\times\\ \Big[&\exp\Big(-\frac{\varepsilon\sqrt{\vartheta\Re'\mathcal{D}'}}{av}x-\sqrt{\vartheta}A(y)\Big)\operatorname{erfc}\Big(\frac{\varepsilon\sqrt{\Re'\mathcal{D}'}}{2av\sqrt{t-x\Re/v}}x+\frac{A(y)}{2\sqrt{t-x\Re/v}}-\sqrt{\vartheta}\sqrt{t-x\Re/v}\Big)+\\ &\exp\Big(\frac{\varepsilon\sqrt{\vartheta\Re'\mathcal{D}'}}{av}x+\sqrt{\vartheta}A(y)\Big)\operatorname{erfc}\Big(\frac{\varepsilon\sqrt{\Re'\mathcal{D}'}}{2av\sqrt{t-x\Re/v}}x+\frac{A(y)}{2\sqrt{t-x\Re/v}}+\sqrt{\vartheta}\sqrt{t-x\Re/v}\Big)\Big]\\ &\text{if}\ \ (t-x\Re/v)>0\\ &\text{otherwise}\ \ \frac{C'}{C_D}=0\quad\text{if}\quad(t-x\Re/v)\le 0\end{aligned}\tag{14.40}$$

where

$$A(y)=\sqrt{\frac{\Re'}{D'}}\,(y-a)\tag{14.41}$$

Note that for evaluating the analytical exp(.)erfc(.) expressions appearing in (14.39) and (14.40) the more suitable exf(., .) function is used as already introduced in Sect. 12.5.1.

To predict the ultimate penetration distances steady-state solutions can be useful. Closed analytical steady-state solutions can be found [506] without the need for

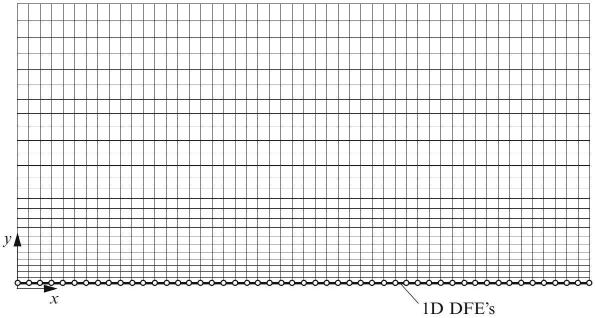


Fig. 14.7 Used nonuniform finite element mesh (vertical exaggeration 300:1) of the half-space fracture matrix domain consisting of 50 × 25 2D quadrilateral porous matrix elements combined with 50 1D DFE's located at $y = a$

neglecting dispersion within the fracture so as necessary for the transient solutions (14.39) and (14.40). The steady-state solute distribution within the fracture results as

$$\frac{C}{C_D} = \exp\left[\left(\frac{v}{2\mathcal{D}} - \sqrt{\frac{v^2}{4\mathcal{D}^2} + \frac{\vartheta + \varepsilon\frac{\sqrt{\mathcal{D}'\vartheta}}{a}}{\mathcal{D}}}\right)x\right] \tag{14.42}$$

and the steady-state solute distribution within the porous matrix results as

$$\frac{C'}{C_D} = \exp\left[\left(\frac{v}{2\mathcal{D}} - \sqrt{\frac{v^2}{4\mathcal{D}^2} + \frac{\vartheta + \varepsilon\frac{\sqrt{\mathcal{D}'\vartheta}}{a}}{\mathcal{D}}}\right)x\right]\exp\left[-\sqrt{\frac{\vartheta}{\mathcal{D}'}}\,(y-a)\right] \tag{14.43}$$

Equation (14.42) can be used to estimate the penetration depth d_δ into the fracture at steady state for a given concentration of $\delta = C/C_D$. It gives

$$d_\delta = \frac{\ln\delta}{\frac{v}{2\mathcal{D}} - \sqrt{\frac{v^2}{4\mathcal{D}^2} + \frac{\vartheta+\varepsilon\frac{\sqrt{\mathcal{D}'\vartheta}}{a}}{\mathcal{D}}}} \tag{14.44}$$

We compare the analytical solutions given by (14.39) for the solute behavior in the fracture and by (14.40) for the solute behavior in the porous matrix with FEFLOW's finite-element simulations based on the spatial discretization shown in Fig. 14.7. The symmetric half of the fracture-matrix domain is discretized by only 50×25 quadrilaterals in variable thicknesses in $y-$direction. The fracture is modeled by using 50 1D DFE's sharing the corresponding quadrilateral element edges of the porous matrix at $y = a$, $0 \leq x \leq L$ (Fig. 14.7). Note that the discretized 2D domain measures $L\times(D-a)$ in $x-$ and $y-$direction (Fig. 14.6) while the thickness (aperture) of the fracture is integrated in the 1D parameters of the used DFE's.

The parameters and conditions used in the numerical simulations are summarized in Table 14.1. Unspecified BC's for flow and solute transport represent boundaries

Table 14.1 Parameters and conditions used for the fractured media diffusion problem

Quantity	Symbol	Value	Unit
Half-domain shown in Fig. 14.6			
Domain length	L	3	m
Domain width	D	$5 \cdot 10^{-3}$	m
Half of fracture width	a	$6 \cdot 10^{-5}$	m
Porous matrix			
Isotropic hydraulic conductivity	K	10^{-22}	$\mathrm{m\,s^{-1}}$
Porosity	ε	0.35	1
Molecular diffusion	$\mathcal{D}'$	$[10^{-10} - 10^{-14}]$	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	0	m
Transverse dispersivity	β_T	0	m
Retardation	$\mathfrak{R}'$	1	1
Decay rate	ϑ	0	$\mathrm{s^{-1}}$
Fracture			
Flow law	Hagen-Poiseuille (Table 4.5, case PN, 1D)		
Hydraulic aperture	$b = 2a$	$1.2 \cdot 10^{-4}$	m
Cross-section area	aB	$6 \cdot 10^{-5}$	$\mathrm{m^2}$
Hydraulic radius	$r_{\mathrm{hydr}} = \frac{b}{2}$	$6 \cdot 10^{-5}$	m
Parameter factor (standard)	$f_0 = \rho_0 g/\mu_0$	$7.55 \cdot 10^6$	$\mathrm{m^{-1}\,s^{-1}}$
Molecular diffusion	$\mathcal{D}$	0	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	0	m
Retardation	$\mathfrak{R}$	1	1
Decay rate	ϑ	0	$\mathrm{s^{-1}}$
Groundwater velocity (steady-state) in the fracture	v	2	$\mathrm{m\,d^{-1}}$
Flow BC			
Neumann-type BC at fracture inlet ($x = 0, y = a$)	$q_h = -v$	-2	$\mathrm{m\,d^{-1}}$
Dirichlet-type BC at fracture outlet ($x = L, y = a$)	h_D	0	m
Solute IC and BC			
Initial condition (IC) of solute	C_0	0	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC at fracture inlet ($x = 0, y = a$)	C_D	1	$\mathrm{mg\,l^{-1}}$
FEM			
Mesh of 50 × 25 quadrilateral elements with 50 1D DFE's (Fig. 14.7), GFEM (no upwind)			
Initial time step size[a]	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Simulation time period	t_{end}	4	d

[a] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$

at which natural BC's are imposed, i.e., $-(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} = 0$ and $-(\boldsymbol{D} \cdot \nabla C') \cdot \boldsymbol{n} = 0$, respectively. For the fracture a Hagen-Poiseuille law of flow motion is assumed (cf. Sect. 4.3.2.2 and Table 4.5). Since the analytical solution (14.39) is only valid for negligible dispersion within the fracture ($\mathcal{D} \equiv 0$), we also set the dispersion to zero for the DFE's in the numerical approach. To stabilize the numerical simulations we prefer the GLS 1st-order accurate FE/BE predictor-corrector time stepping method, however, no resort to upwinding is necessary.

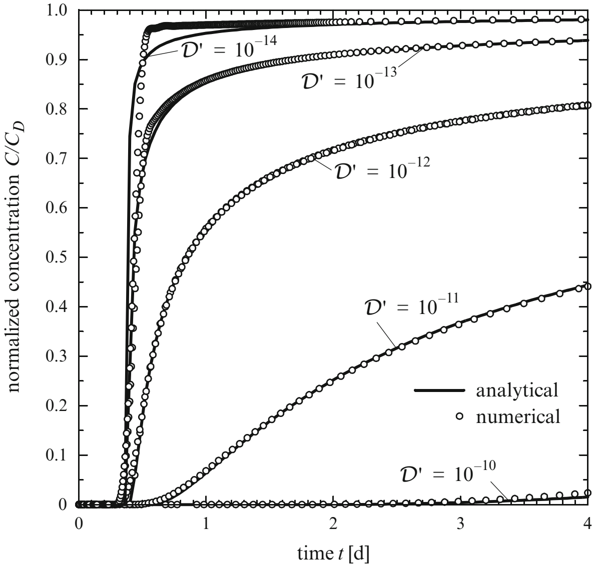


Fig. 14.8 Simulated versus analytical solute breakthrough curves in the fracture ($y = a$) at distance of $x = 0.76$ m from the source point for values of matrix diffusion $\mathcal{D}'$ in the range of 10^{-10}–10^{-14} m^2 s^{-1}

FEFLOW's finite-element results are compared in Figs. 14.8–14.10 with the analytical findings. The agreement is rather well, although the used mesh is relatively coarse. Figure 14.8 shows the solute breakthrough curves in the fracture for values of different matrix diffusion $\mathcal{D}'$. Differences to the analytical solutions are only revealed for very small diffusion, i.e., for cases where the advective solute transport is dominant in the fracture. This is also seen in the computed solute profiles into the porous matrix as depicted in Fig. 14.10, where the used discretization in $y-$direction is obviously insufficient for a small matrix diffusion of $\mathcal{D}' = 10^{-14}$ m^2 s^{-1}. The results for this case can be improved by using more refined meshes.

For the solute profiles in the porous matrix in longitudinal $x-$direction shown in Fig. 14.9 we can observe that the accuracy of the numerical results expectedly decreases with increasing matrix diffusion, such as revealed in particular for $\mathcal{D}' = 10^{-10}$ m^2 s^{-1} in Fig. 14.9. A more refined mesh could also improve the accuracy for those cases of dominant matrix diffusion. The numerical simulations required numbers of adaptive time steps ranging between 113 and 266 for simulating a time period of 4 days in dependence on the used matrix diffusions $\mathcal{D}'$.

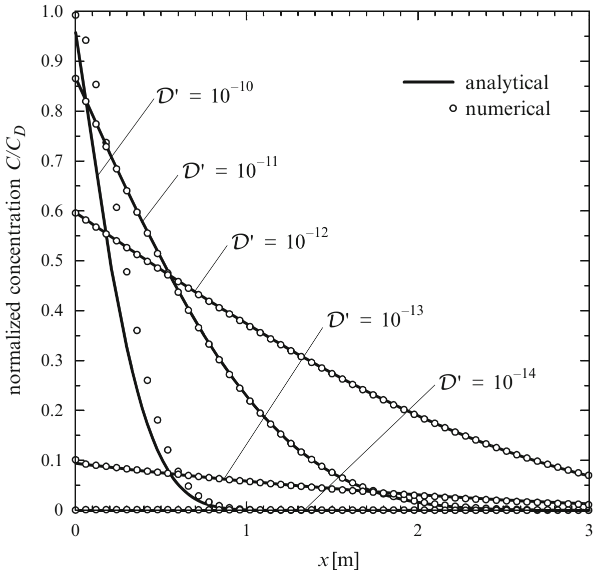


Fig. 14.9 Simulated versus analytical solute profiles at $t = 4$ days along the porous matrix in $x-$direction at $y = 10^{-4}$ m for values of matrix diffusion $\mathcal{D}'$ in the range of 10^{-10}–10^{-14} m^{2}s^{-1}

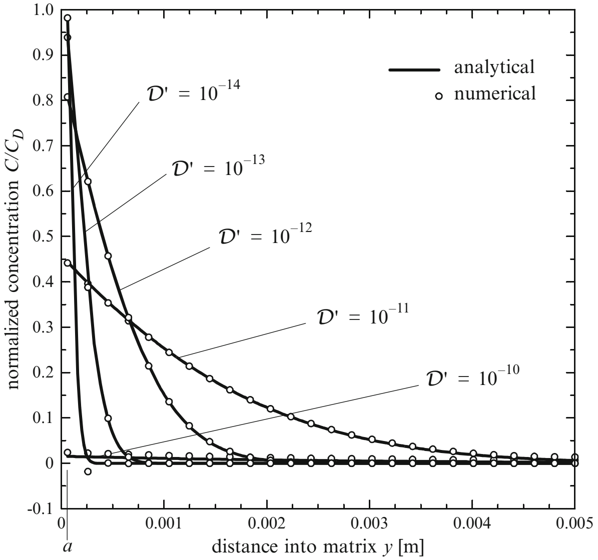


Fig. 14.10 Simulated versus analytical solute profiles at $t = 4$ days into the porous matrix in $y-$direction at $x = 0.76$ m for values of matrix diffusion $\mathcal{D}'$ in the range of 10^{-10}–10^{-14} m^2 s^{-1}

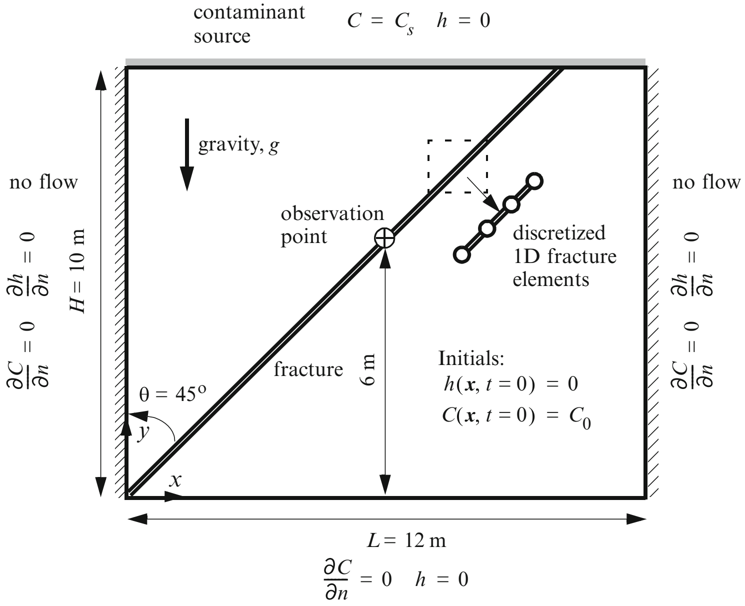

contaminant source
$C = C_s \quad h = 0$
gravity, g
no flow
no flow
$\frac{\partial C}{\partial n} = 0 \quad \frac{\partial h}{\partial n} = 0$
$\frac{\partial C}{\partial n} = 0 \quad \frac{\partial h}{\partial n} = 0$
$H = 10$ m
observation point
discretized 1D fracture elements
fracture
6 m
$\theta = 45°$
y
x
Initials:
$h(\boldsymbol{x}, t = 0) = 0$
$C(\boldsymbol{x}, t = 0) = C_0$
$L = 12$ m
$\frac{\partial C}{\partial n} = 0 \quad h = 0$

Fig. 14.11 Single 45°−inclined fracture in a porous matrix; 2D geometry, BC's and IC's (Modified from [198])

14.5.2 Density-Dependent Solute Transport in a 45°−Inclined Single Fracture Embedded in a Low-Permeable Porous Matrix

Graf and Therrien [198] have studied density-dependent solute transport in single fractures of arbitrary inclination embedded in a low-permeable porous matrix. We shall benchmark their results for the 45°−inclined fracture problem against FEFLOW and the research code Ground Water (GW) developed by F. Cornaton [102]. This single fracture problem is shown in Fig. 14.11. The fracture inclined by 45° is discretized by using correspondingly inclined 1D DFE's. The left and right boundaries of the $L \times H$ enclosing porous matrix domain are assumed to be impermeable. The top and bottom boundaries are modeled as open boundaries with a constant hydraulic head h (set to zero). A contaminant source of constant solute concentration $C = C_s$ overlies groundwater of initial concentration $C = C_0$, where $C_0 = 0 < C_s = 1$. The simulations cover a time of 20 years. The model parameters and conditions are summarized in Table 14.2. It is assumed that the porous matrix is isotropic and homogenous and that the entire domain is completely saturated. BC's unreported in Table 14.2 for flow and solute transport represent boundaries at which natural BC's are imposed, i.e., $-(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} = 0$ and $-(\boldsymbol{D} \cdot \nabla C) \cdot \boldsymbol{n} = 0$, respectively.

Table 14.2 Parameters and conditions used for the inclined single fracture problem

Quantity	Symbol	Value	Unit
Study domain shown in Fig. 14.11.			
Domain length	L	12	m
Domain height	H	10	m
Fracture slope	θ	45	°
Fracture aperture	b	$5 \cdot 10^{-5}$	m
Porous matrix			
Isotropic hydraulic conductivity	K	$8.7216 \cdot 10^{-9}$	$\mathrm{m\,s^{-1}}$
Specific storage coefficient	S_o	$1.743 \cdot 10^{-5}$	$\mathrm{m^{-1}}$
Specific solutal expansion coefficient	α	0.2	1
Porosity	ε	0.35	1
Molecular diffusion coefficient	D	$5 \cdot 10^{-10}$	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	0.1	m
Transverse dispersivity	β_T	0.005	m
Fracture			
Flow law	Hagen-Poiseuille (Table 4.5, case PN, 1D)		
Fracture area	bB	$5 \cdot 10^{-5}$	$\mathrm{m^2}$
Corrected hydraulic aperture[a]	b_{corr}	$5.374 \cdot 10^{-5}$	m
Effective hydraulic radius	$r_{\mathrm{hydr}} = \frac{b_{\mathrm{corr}}}{2}$	$2.687 \cdot 10^{-5}$	m
Specific storage coefficient	S_o	$4.4 \cdot 10^{-6}$	$\mathrm{m^{-1}}$
Specific solutal expansion coefficient	α	0.2	1
Molecular diffusion coefficient	D	$5 \cdot 10^{-9}$	$\mathrm{m^2\,s^{-1}}$
Longitudinal dispersivity	β_L	0.1	m
Flow IC and BC's			
Initial condition (IC) of hydraulic head	h_0	0	m
Dirichlet-type BC on top at $(0 \leq x \leq L, y = H)$	h_D	0	m
Dirichlet-type BC on bottom at $(0 \leq x \leq L, y = 0)$	h_D	0	m
Solute IC and BC			
Initial condition (IC) of solute	C_0	0	$\mathrm{mg\,l^{-1}}$
Dirichlet-type BC on top at $(0 \leq x \leq L, y = H)$	C_D	1	$\mathrm{mg\,l^{-1}}$
FEM			
Uniform mesh of 24,000 triangles with 100 1D inclined DFE's (Fig. 14.12), GFEM (no upwind)			
Constant fully implicit time stepping variant 1 (combined with Picard iteration):			
Constant time step size	Δt	0.2	years
Adaptive time stepping variant 2 (AB/TR predictor-corrector strategy):			
Initial time step size[b]	Δt_0	10^{-5}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	20	years

[a] Hydraulic aperture b has to be corrected by the factor $\sqrt{f/f_0}$ due to a different viscosity magnitude, where $f = \rho g/\mu$ and $f_0 = \rho_0 g/\mu_0 = 7.55 \cdot 10^6\ \mathrm{m^{-1}s^{-1}}$, see Sect. 4.4.2

[b] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$ and maximum time step size $\Delta t_{\max} = 73$ d

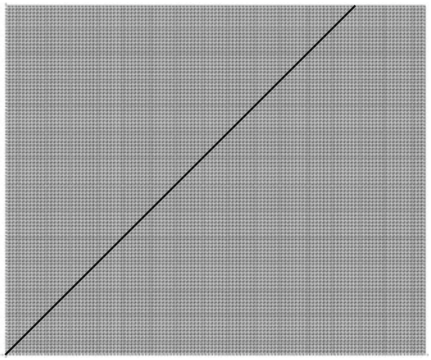


Fig. 14.12 2D triangular finite element mesh with 1D DFE's used for FEFLOW and GW simulations

Graf and Therrien [198] tested different fracture slopes θ and mesh refinement levels. The present study focuses on the 45°−inclined fracture problem at their highest grid refinement level, consisting of 12,221 nodes and 24,000 triangles as shown in Fig. 14.12. We use two time stepping strategies: (1) in agreement to Graf and Therrien [198] a fully implicit time step marching scheme (combined with a Picard iteration) with a constant time step length Δt of 0.2 years, and (2) alternatively, the adaptive GLS 2nd-order accurate predictor-corrector AB/TR time stepping using a RMS tolerance error of 10^{-4}. No upwinding is employed in all simulations. The computation of the consistent velocity fields is performed by using FKA. The inclined fracture is modeled by 100 1D DFE's fitted to the edges of the corresponding triangular elements (Fig. 14.12). For the flow in the fracture the Hagen-Poiseuille law is applied. Fluid viscosity is considered independent of the concentration $\mu = \mu_0 =$ const. Graf and Therrien's variable-density computations employed standard Oberbeck-Boussinesq (OB) approximation (cf. Sect. 3.10.3).

For the 45°−inclined fracture problem the results obtained by Graf and Therrien [198] and by FEFLOW in form of computed concentration distributions as well as velocity fields and pathline patterns at 2, 4 and 10 years simulation time are shown in Fig. 14.13. It reveals how the solutes migrate from the fracture into the adjoining porous matrix mainly governed by hydrodynamic dispersion and to a small degree by convection. As a typical feature of the problem two convection cells form above and below the fracture with increasing extent in time. Both cells move downward in time. Note that the cell above the fracture moves faster downward than the lower cell. Both convection cells remain separated by the high-conductive fracture, therefore, acts as a barrier to convection.

At a first glance, FEFLOW and Graf and Therrien's results agree very well. However, as already seen in Fig. 14.13 the advance of solute transport in the fracture seems slightly faster at early times in Graf and Therrien's predictions compared to the FEFLOW results. Indeed, this can be confirmed if as shown in Fig. 14.14. While the FEFLOW curves for adaptive time stepping (taking 236 steps) and for constant time steps (100 implicit steps with each of 0.2 years length) provide

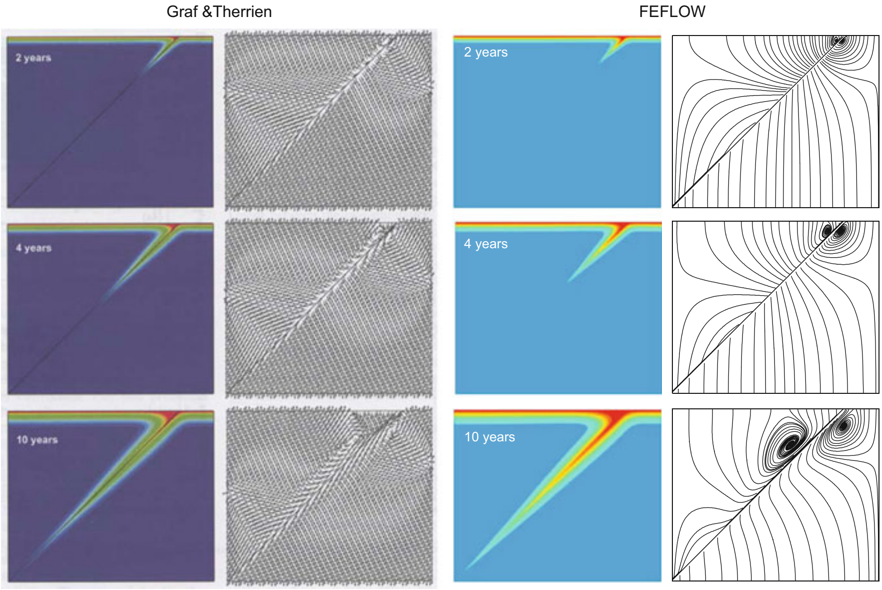


Fig. 14.13 Computed concentration distributions and velocity/pathline field after 2, 4 and 10 years simulation. Comparison of FEFLOW results obtained by using AB/TR time stepping (*right*) to findings by Graf and Therrien [198] modeled by a fully implicit constant time stepping (*left*). OB approximation is used

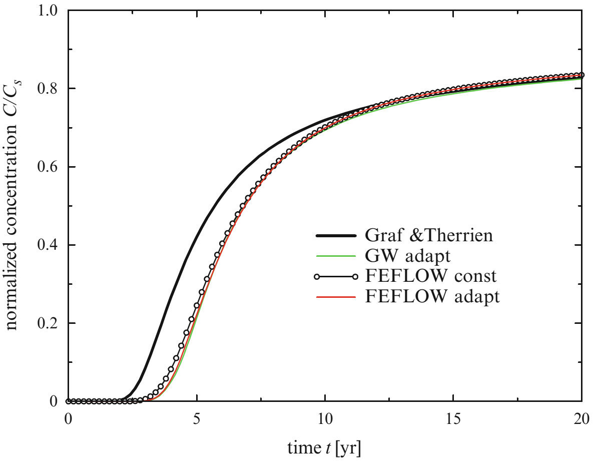


Fig. 14.14 Breakthrough curves at observation point $x = y = 6\,\text{m}$ (Fig. 14.11). Comparison of Graf and Therrien's results [198] to GW [102] (with adaptive time stepping) and FEFLOW (with constant and adaptive time stepping) in using OB approximation

reasonably close solutions, Graf and Therrien's breakthrough curve is apparently advanced at early times. Due to the high velocity contrasts between matrix and

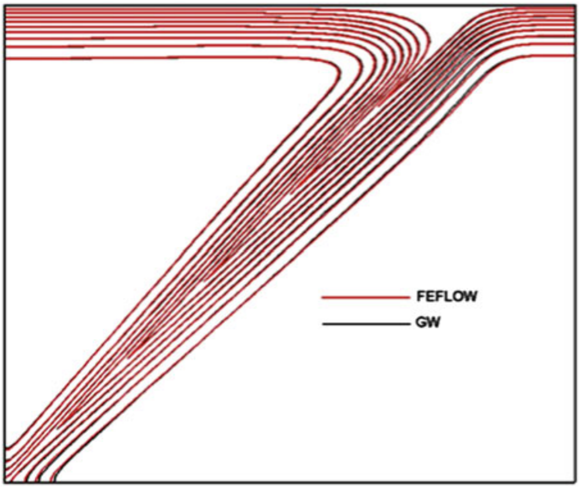


Fig. 14.15 Computed solute concentration contours at $t = 15$ years: FEFLOW versus GW results in using OB approximation

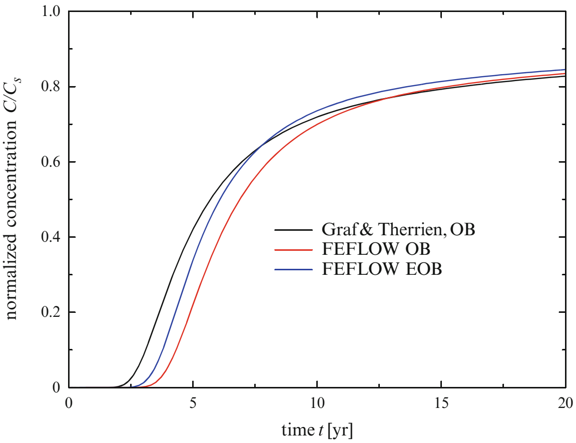


Fig. 14.16 Breakthrough curves at the observation point $x = y = 6\,\text{m}$ (Fig. 14.11). Comparison between OB approximation and EOB approximation. Adaptive time stepping is used for FEFLOW's solutions

fracture, the influence of early times on the spreading of solute in the depth is crucial and requires further model comparisons.

The problem was also simulated by using the GW finite-element simulator [102]. The GW results provide a nearly perfect agreement with the FEFLOW predictions (cf. Figs. 14.14–14.16). As evidenced in Fig. 14.14 FEFLOW's and GW's breakthrough curves are very close. This could be confirmed by using both adaptive and constant time stepping strategies. Note further that the type of solving the resulting sparse equation systems did not influence the outcome. Direct and iterative equation solvers were tested in FEFLOW. Additionally, the extended Boussinesq approximation (EOB), cf. Sect. 3.10.3, is also performed. As indicated in Fig.14.16 the breakthrough curve for the EOB is slightly shifted in advance compared to FEFLOW's OB solution, however, remains further behind Graf and Therrien's OB solution. It can be concluded that the discrepancies between Graf

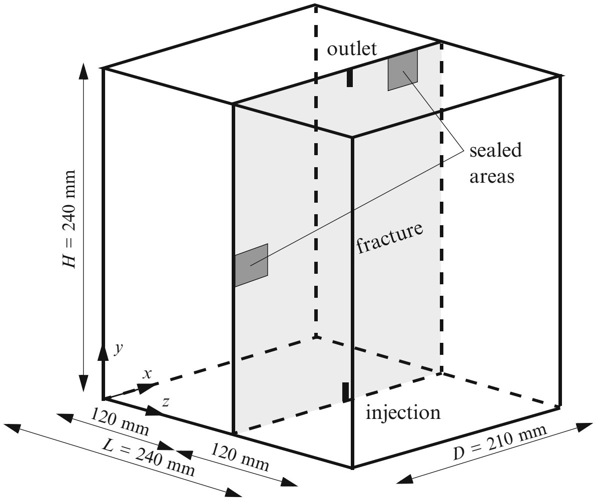


Fig. 14.17 Schematic representation of the fractured sandstone block

and Therrien's findings and the results simulated by FEFLOW or GW are not attributed to different time stepping strategies, Boussinesq approximations and different sparse matrix solvers. Furthermore, more spatially refined meshes did not change notably anymore the solutions because the mesh convergence is practically achieved at the analyzed mesh refinement level.

14.5.3 Wendland and Himmelsbach's Experiment: Solute Transport in a 3D Fracture-Matrix System

Wendland and Himmelsbach [560] conducted laboratory experiments and numerical computations of solute transport in a fractured sandstone block. The sandstone block has a length of 24 cm, a width of 21 cm and a height of 24 cm (Fig. 14.17). The fracture with a mean aperture of $>350\,\mu$m divides the block into two parts. The geometric details of the fracture plane are shown in Fig. 14.18. Water is pumped from below flowing upwards at a constant rate of $Q = 4.57\,\mathrm{ml\,h^{-1}}$ through the fracture plane. A multi-tracer experiment with pyranine and cadmium as solutes was performed. The tracer is injected in the fracture at the bottom and the tracer breakthrough is observed at the outlet on the top of the fracture (Figs. 14.17 and 14.18).

In the tracer experiment the injection of the solutes is considered as a pulse directly into the fracture. The injection of the total tracer mass of 32.2 μg lasted less than 1 min. For modeling purposes, Wendland and Himmelsbach [560] smoothed the pulse injection over a time interval of 6 min, which is still small relative to the duration of the tracer experiment. Geometric relations, IC, BC's and material parameters used for the present simulation are summarized in Table 14.3. BC's unreported in Table 14.3 for flow and solute transport represent boundaries at which

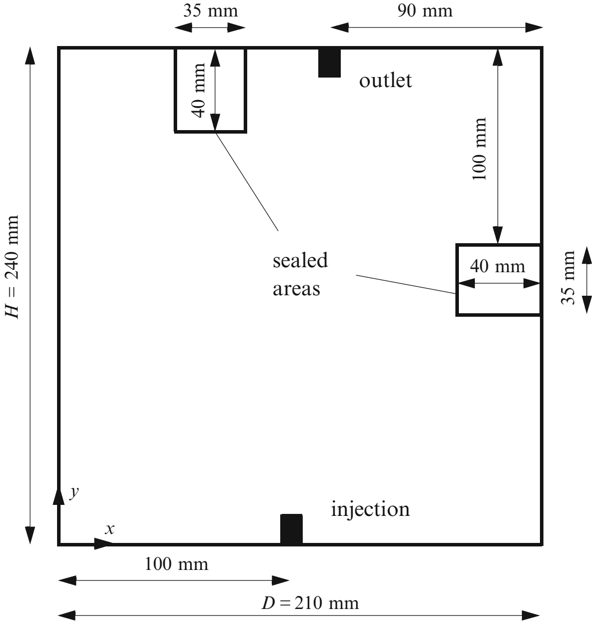


Fig. 14.18 Geometry of the fracture plane at $z = L/2 = 120\,\text{mm}$

natural BC's are imposed, i.e., $-(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} = 0$ and $-(\boldsymbol{D} \cdot \nabla C) \cdot \boldsymbol{n} = 0$, respectively. The flow is modeled steady-state while the solute transport is transient.

In Wendland and Himmelsbach's simulation [560], the sandstone block was discretized into 8,668 3D elements for the porous matrix and 435 2D elements for the plane fracture. They used a symmetric streamline stabilization technique (cf. Sect. 8.14.5) and an implicit time stepping with 2,000 constant time steps. In the present FEFLOW simulation the spatial discretization is largely similar to Wendland and Himmelsbach's mesh in that, considering the expected concentration profile in the porous matrix, logarithmic grid spacing is employed. The first nodal row is located at a distance of $2 \cdot 10^{-5}$ m from and parallel to the fracture interface. The subsequent nodes are at distances of $5 \cdot 10^{-5}$, $1.4 \cdot 10^{-4}$ and $4 \cdot 10^{-3}$ m. All further nodes parallel to the vertical fracture are located at a constant horizontal distance of 1 cm, except the last two slices having distances of 4 cm. The resulting finite element mesh of the entire sandstone block is shown in Fig. 14.19. Note that in the FEFLOW simulations only the symmetric half of the domain is considered. This leads to a half-mesh consisting of 20,160 3D linear brick elements for the porous matrix and 1,904 2D linear quadrilateral fracture elements. In order to account for the sealed areas in the fracture plane (Fig. 14.18), the corresponding 2D elements of the fracture were deleted from the mesh.

In the FEFLOW simulation the adaptive GLS FE/BE predictor-corrector scheme is applied with an initial time step of 10^{-5} d and a RMS error criterion of 10^{-4}. The simulations are performed for a period of 660 min, which required 173 variable

Table 14.3 Parameters and conditions used for Wendland and Himmelsbach's 3D fracture-matrix problem

Quantity	Symbol	Value	Unit
Domain and fracture plane shown in Figs. 14.17 and 14.18, respectively.			
Domain width	L	240	mm
Domain depth	D	210	mm
Domain height	H	240	mm
Flux at injection point	Q	4.57	ml h^{-1}
Porous matrix			
Isotropic hydraulic conductivity	K	$1 \cdot 10^{-9}$	m s^{-1}
Porosity	ε	0.085	1
Molecular diffusion coefficient	D	$5 \cdot 10^{-11}$	$\text{m}^2\,\text{s}^{-1}$
Longitudinal dispersivity	β_L	0	m
Transverse dispersivity	β_T	0	m
Fracture			
Flow law	Hagen-Poiseuille (Table 4.5, case PN, 2D)		
Fracture aperture	b	507	µm
Fracture thickness of half-space	B	$2.535 \cdot 10^{-4}$	m
Effective hydraulic radius	$r_{\text{hydr}} = \frac{b}{2}$	$2.535 \cdot 10^{-4}$	m
Parameter factor (standard)	$f_0 = \rho_0 g / \mu_0$	$7.55 \cdot 10^{6}$	$\text{m}^{-1}\,\text{s}^{-1}$
Molecular diffusion coefficient	D	0	$\text{m}^2\,\text{s}^{-1}$
Longitudinal dispersivity	β_L	0.006	m
Transverse dispersivity	β_T	0	m
Steady-state flow BC's			
Dirichlet-type BC at outlet ($x = 120$ mm, $y = H, z = L/2$)	h_D	0	m
Well-type SPC at inlet, discharge in the half-space ($x = 100$ mm, $y = 0, z = L/2$)	$Q_w = -\frac{Q}{2}$	$-5.484 \cdot 10^{-5}$	$\text{m}^3\,\text{d}^{-1}$
Solute IC and BC			
Initial condition (IC) of solute	C_0	0	mg l^{-1}
Dirichlet-type BC at inlet ($x = 100$ mm, $y = 0, z = L/2$)	C_D	$\begin{cases} 70.46 & \text{at } t \leq 6 \text{ min} \\ 0 & \text{at } t > 6 \text{ min} \end{cases}$	mg l^{-1}
FEM			
Nonuniform 42 × 48 × 10 mesh of 20,160 brick elements in variable extents for the simulation half-domain ($0 \leq x \leq D,\ 0 \leq y \leq H,\ 0 \leq z \leq L/2$), PGLS upwinding			
Initial time step size	Δt_0	10^{-5}	d
RMS error tolerance (FE/BE)	ϵ	10^{-4}	1
Simulation time period	t_{end}	660	min

time steps. Similar to Wendland and Himmelsbach [560] a PGLS upwind technique (cf. Sect. 8.14.5) is used to stabilize the numerical solution.

Figure 14.20 illustrates the resulting flow field and head distribution in the fracture-matrix system. The computed distribution of solutes within the fracture at the final time of 660 min is shown in Fig. 14.21. Wendland and Himmelsbach's [560] results are displayed in Fig. 14.22. A comparison of the FEFLOW results with solution given by Wendland and Himmelsbach [560] reveals differences in the solute concentration at the sealed areas and near the outlet of the fracture. We note,

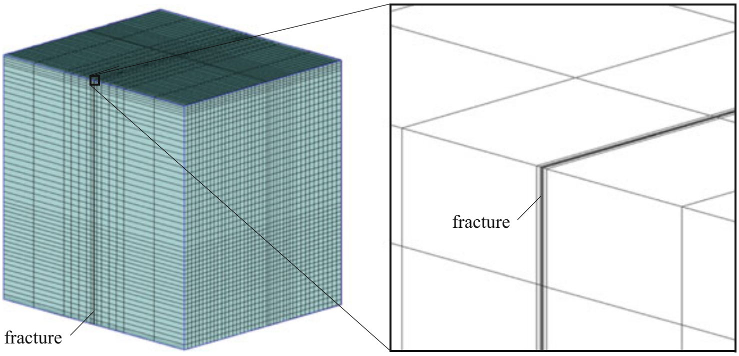


Fig. 14.19 FEFLOW's finite element mesh of the sandstone block with a vertical fracture: view of the entire block and magnified mesh at the fracture

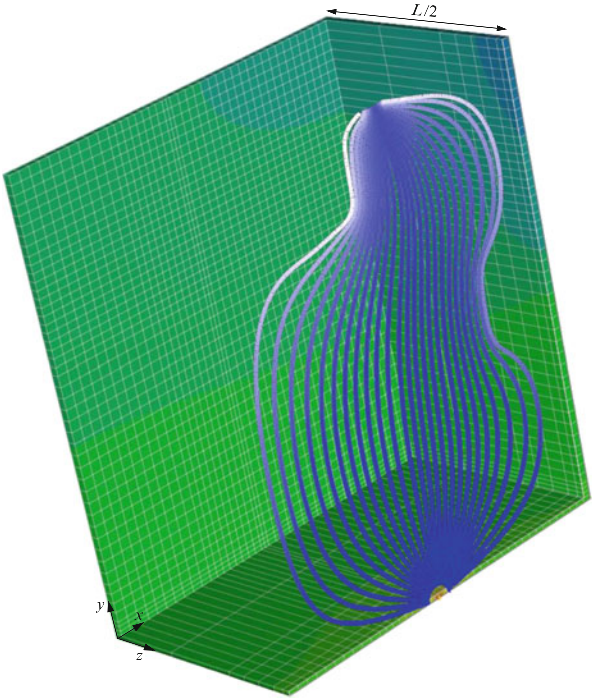


Fig. 14.20 Computed stationary pathlines in the fracture and head distributions in the contacted sandstone (half-space view)

however, that the magnitudes of solute concentrations are in good agreement (the same concentration levels are used both in Figs. 14.21 and 14.22). Perhaps more significant are the results of the breakthrough behavior at the outlet shown in Figs. 14.23 and 14.24. The agreement with Wendland and Himmelsbach's [560] measurements is quite well. Wendland and Himmelsbach obtained a higher peak

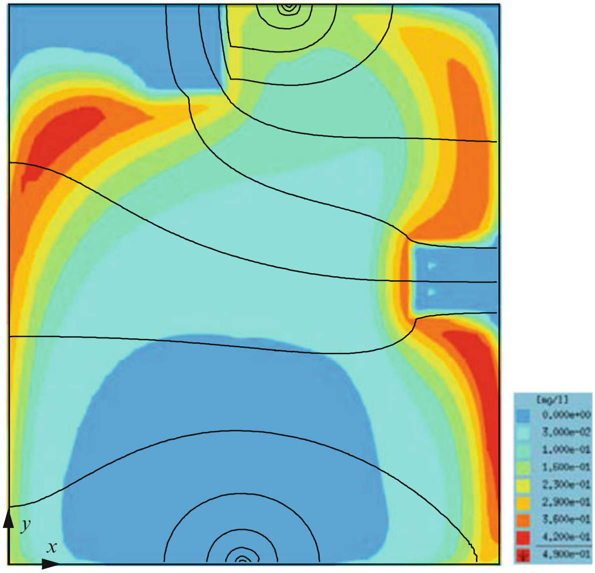


Fig. 14.21 FEFLOW results of solute distribution within the fracture ($z = L/2$) at final simulation time $t = 660\,\text{min}$

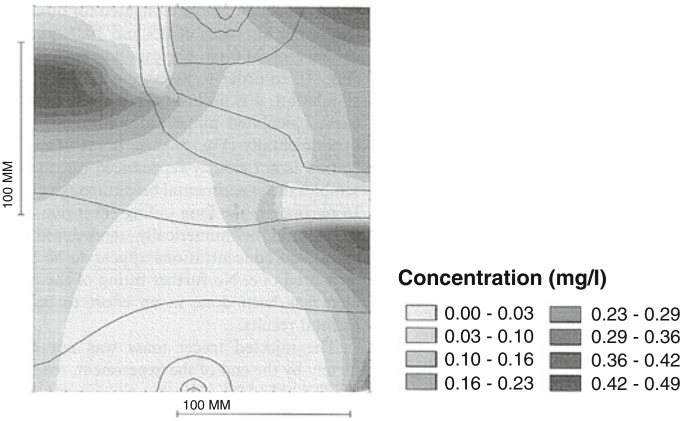


Fig. 14.22 Wendland and Himmelsbach's [560] simulation results of solute distribution within the fracture at final simulation time $t = 660\,\text{min}$

concentration in their simulations compared to the measurements (Fig. 14.24) and the FEFLOW simulation (Fig. 14.23). Obviously, the solute diffusion into the matrix and its accurate numerical representation in the 3D fracture-matrix system is of high importance. The better agreement of the FEFLOW results can result from the more refined spatial resolution.

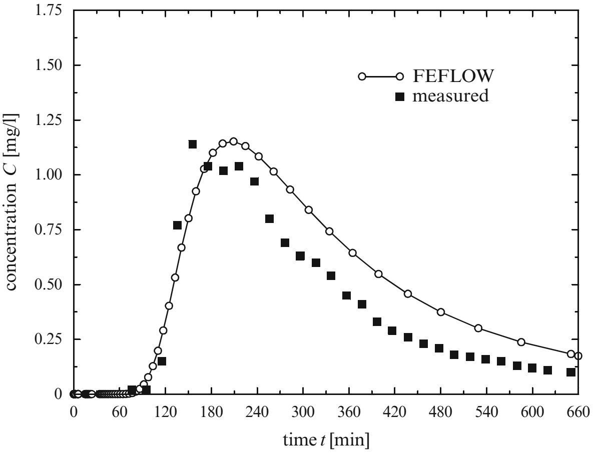


Fig. 14.23 Breakthrough curve at the outlet: FEFLOW results compared to the measurements [560]

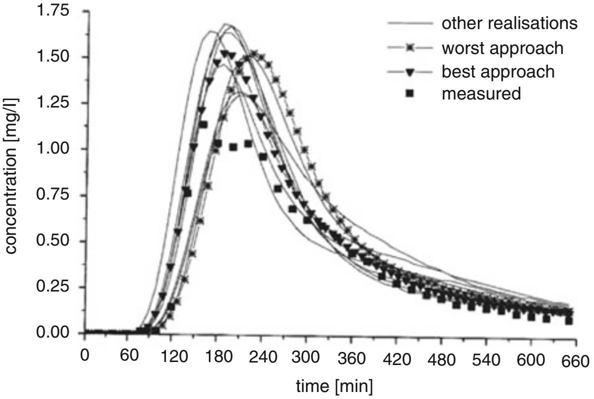


Fig. 14.24 Measured and simulated breakthrough curves at the outlet obtained by Wendland and Himmelsbach [560]

14.5.4 Flow and Solute Transport in a Fracture Network of Rock Mass

The simulation of flow and transport processes in a collection of individual fractures (fracture network) is a challenging task due to the inherent geometric complexity and its required numerical resolution.[3] While a small set of individual fractures

[3] I acknowledge F. Cornaton (DHI-WASY) for providing the fracture network generation and simulation results performed by the finite-element simulator Ground Water (GW) [102].

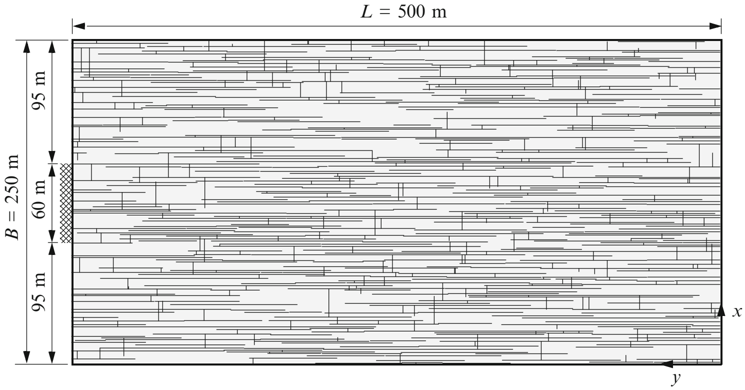


Fig. 14.25 Study domain and fracture network generated by Josnin et al.'s algorithm [293] using shape parameters $E = F = 0.5\,\mathrm{m}$. At central LHS boundary ($95\ \mathrm{m} \leq x \leq 155\ \mathrm{m}, y = L$) a solute source is imposed

can still often be described in a deterministic way, a fracture network, where a whole set of crossing and intersecting fractures is typical, necessitates more advanced modeling approaches [5]. Fracture networks are usually described either via stochastic or fractal approaches [80] or by using mechanical parameters in combination with statistical rules for the underlain rock masses [293, 294].

We consider a 2D example of a sedimentary rock mass measuring $B \times L = 250 \times 500\,\mathrm{m}$ (Fig. 14.25). A fracture network is generated by using the algorithm developed by Josnin et al. [293] based on stochastic and mechanical parameters given for a tabular stratified rock. A discontinuity network results which is composed of two orthogonal joint sets normal to bedding in the tabular sedimentary rock mass, controlled by two shape parameters: the half-wide E and parameter F for adjusting joint overlap. The resulting orthogonal fracture network shown in Fig. 14.25 was generated by choosing $E = F = 0.5\,\mathrm{m}$.

The fracture network geometry (Fig. 14.25) is mapped onto a regular finite-element mesh consisting of 305×984 linear quadrilateral elements. The individual fractures are assigned to the edges of corresponding quadrilaterals. In doing so, 68,488 1D DFE's finally result to model the fracture network in the spatially discretized domain. For the fracture network the Hagen-Poiseuille law of flow with uniform apertures of $100\,\mu\mathrm{m}$ is assumed. A steady-state flow is modeled by prescribing a hydraulic gradient of $1\,\%$ between the LHS boundary at $y = L = 500\,\mathrm{m}$ and the RHS boundary at $y = 0$, ($0 \leq x \leq B = L/2 = 250\,\mathrm{m}$). At the central LHS boundary a single-species solute intrudes into the domain with a constant concentration C_D, migrates through the fracture network and penetrates the porous matrix over time. The simulations cover a time of 1,000 years. The model parameters and conditions are summarized in Table 14.4. It is assumed that the porous matrix is isotropic and homogenous and that the entire domain is completely saturated. BC's unreported in Table 14.4 for flow and solute transport

Table 14.4 Parameters and conditions used for the fracture network model problem

Quantity	Symbol	Value	Unit
Domain and fracture network shown in Fig. 14.25.			
Domain length	L	500	m
Domain width	$B = \frac{L}{2}$	250	m
Porous matrix			
Isotropic hydraulic conductivity	K	$1 \cdot 10^{-8}$	m s^{-1}
Porosity	ε	0.13	1
Molecular diffusion coefficient	D	$5 \cdot 10^{-10}$	$\text{m}^2\,\text{s}^{-1}$
Longitudinal dispersivity	β_L	0	m
Transverse dispersivity	β_T	0	m
Fracture network			
Flow law	Hagen-Poiseuille (Table 4.5, case PN, 1D)		
Fracture area	A	$1 \cdot 10^{-4}$	m^2
Fracture aperture	b	$1 \cdot 10^{-4}$	m
Effective hydraulic radius	$r_{\text{hydr}} = \frac{b}{2}$	$5 \cdot 10^{-5}$	m
Parameter factor (standard)	$f_0 = \rho_0 g / \mu_0$	$7.55 \cdot 10^{6}$	$\text{m}^{-1}\text{s}^{-1}$
Molecular diffusion coefficient	D	$5 \cdot 10^{-9}$	$\text{m}^2\,\text{s}^{-1}$
Longitudinal dispersivity	β_L	0.1	m
Transverse dispersivity	β_T	0	m
Steady-state flow BC's			
Dirichlet-type BC at LHS $(0 \leq x \leq B, y = L)$	h_{D_1}	5	m
Dirichlet-type BC at RHS $(0 \leq x \leq B, y = 0)$	h_{D_2}	0	m
Solute IC and BC's			
Initial condition (IC) of solute	C_0	0	mg l^{-1}
Dirichlet-type BC at central LHS $(95\text{ m} \leq x \leq 155\text{ m}, y = L)$	C_D	1	mg l^{-1}
Dirichlet-type BC at remaining LHS $(0 \leq x < 95\text{ m}, y = L)$ and $(155\text{ m} < x \leq B, y = L)$	C_{D_0}	0	mg l^{-1}
FEM			
Uniform 305 × 984 mesh of 300,120 linear quadrilateral elements with 68,488 1D DFE's (Fig. 14.25), GFEM (no upwind)			
Initial time step size[a]	Δt_0	10^{-3}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	1,000	years

[a] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$ and maximum time step size $\Delta t_{\max} = 50$ years

represent boundaries at which natural BC's are imposed, i.e., $-(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} = 0$ and $-(\boldsymbol{D} \cdot \nabla C) \cdot \boldsymbol{n} = 0$, respectively.

In the FEFLOW simulations the GFEM (without any upwind) and the adaptive GLS 2nd-order accurate predictor-corrector AB/TR time integrator with a RMS tolerance error of 10^{-4} are used. To evaluate the computational results, comparisons to the finite-element research code Ground Water (GW) [102] are performed. GW

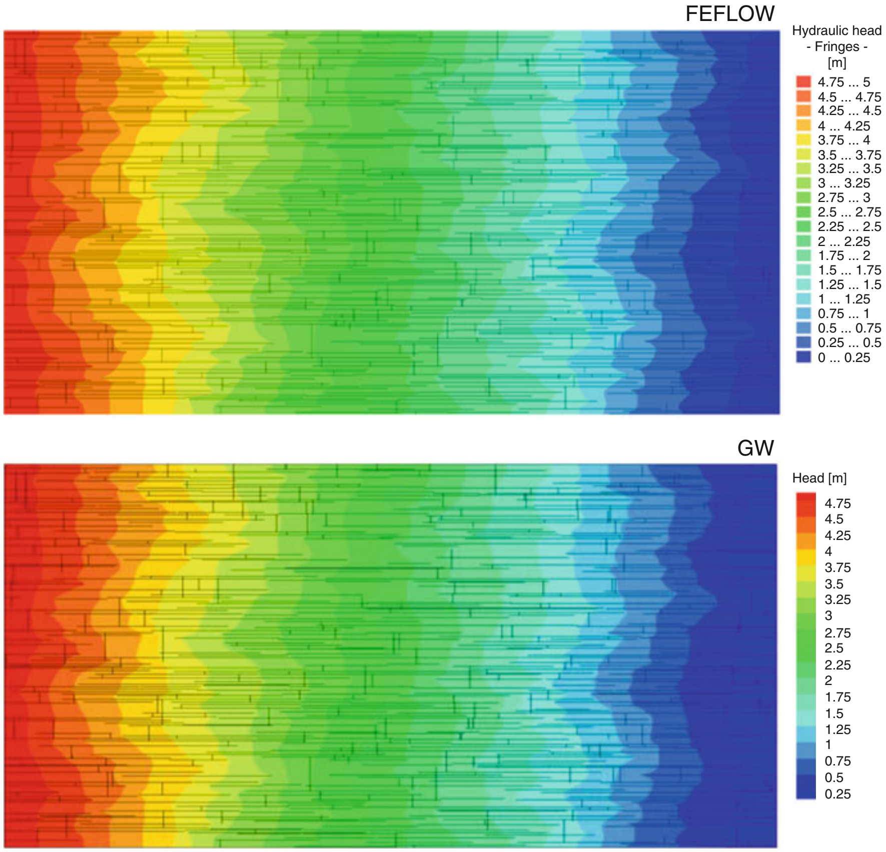


Fig. 14.26 Steady-state hydraulic head distribution $h(\boldsymbol{x})$ in the fracture network domain: FEFLOW vs. GW simulation results

is independently developed and uses differently implemented solution techniques. FEFLOW and GW can run on the same mesh and fracture network data.

The steady-state hydraulic head distribution $h(\boldsymbol{x})$ in the fracture network domain is compared in Fig. 14.26 between FEFLOW and GW revealing a nearly perfect agreement. This can also be evidenced in more detail for $h-$profiles such as exemplified in Fig. 14.27 at $y = 400\,\mathrm{m}$, $0 \leq x \leq B$. For the transient solute transport through the fracture network domain we also recognize very good agreements between FEFLOW's and GW's computational results. This is evidenced in Fig. 14.28 comparing the solute distributions at three selected time stages, in Fig. 14.29 showing the FEFLOW vs. GW solute breakthrough curves at four points selected in the fracture network domain and in Fig. 14.30 comparing $C-$profiles for the cross section at $y = 400\,\mathrm{m}$, $0 \leq x \leq B$. FEFLOW took 226 variable AB/TR time steps for the simulation period of 1,000 years. A different, but likewise variable time stepping of 2nd-order accuracy was used in the GW simulations.

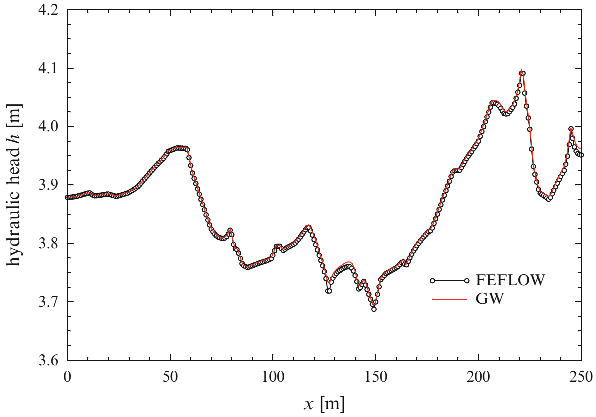


Fig. 14.27 Hydraulic head profiles at $y = 400\,\mathrm{m}$, $0 \le x \le B$, in the fracture network domain simulated by FEFLOW and GW

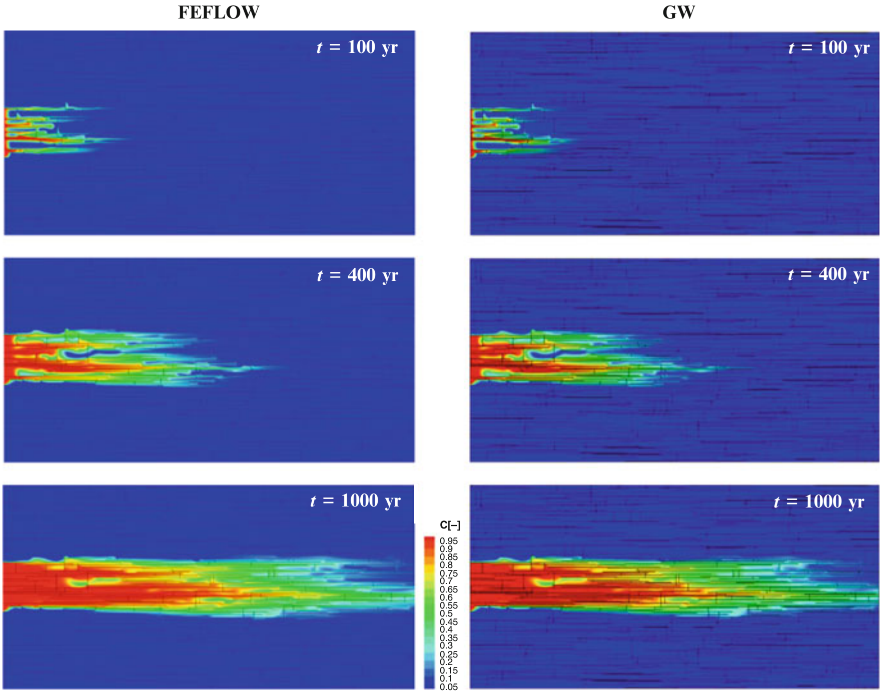


Fig. 14.28 Comparison between FEFLOW's and GW's solute distributions simulated in the fracture network domain at different times *t* (years)

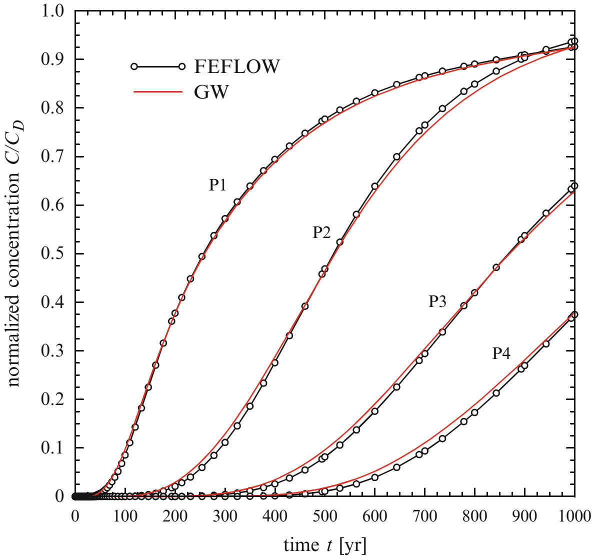


Fig. 14.29 Comparison between FEFLOW's and GW's solute breakthrough curves at points P1$(x, y) = (125\,\text{m}, 400\,\text{m})$, P2$(x, y) = (125\,\text{m}, 300\,\text{m})$, P3$(x, y) = (125\,\text{m}, 200\,\text{m})$ and P4$(x, y) = (125\,\text{m}, 100\,\text{m})$ in the fracture network domain

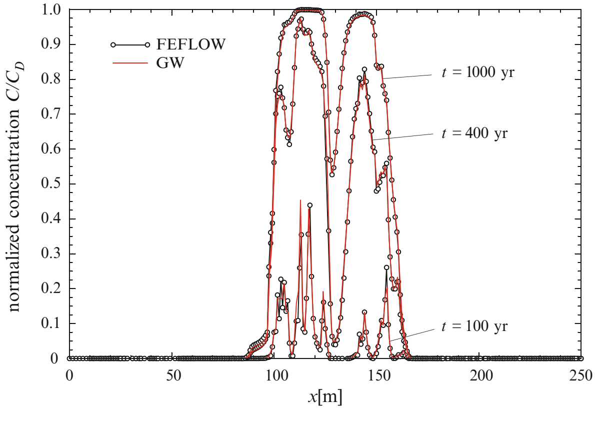


Fig. 14.30 Concentration profiles at $y = 400\,\text{m}$, $0 \leq x \leq B$, in the fracture network domain for different times t (years) simulated by FEFLOW and GW

14.5.5 Thermohaline Variable-Density Convection in an Aquifer-Aquitard-Aquifer System with Abandoned Borehole

In this hypothetical example we study the effect of a single abandoned borehole causing a short-circuit flow situation in a deep stratified aquifer-aquitard-aquifer

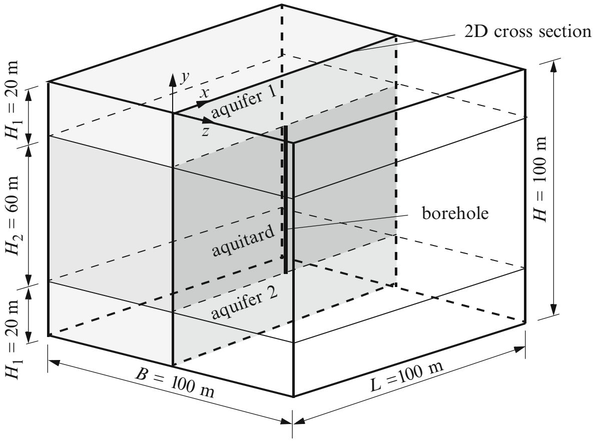


Fig. 14.31 Schematic representation of the aquifer-aquitard-aquifer system with the abandoned borehole in the center of the aquitard

system driven by heavy saltwater and buoyant thermal gradients (Fig. 14.31).[4] The abandoned borehole is to be modeled via the discrete feature approach. The borehole bridges very locally the upper and lower aquifer so that saltwater and heat can be efficiently exchanged over this preferential flow channel. The study domain measures $L \times H \times B = 100 \times 100 \times 100$ m for a 3D schematization and $L \times H = 100 \times 100$m for a 2D cross-sectional schematization as shown in Fig. 14.31. The upper and lower aquifers have thicknesses of each 20 m, the aquitard in between is 60 m thick. In the center of the domain the abandoned borehole is located, which interconnects the upper and the lower aquifer in a vertical distance of 60 m. Traces of the abandoned borehole in the aquifers are neglected.

At initial time, the aquifer system is in a stable hydrostatic equilibrium: the model domain contains freshwater and is subjected to a thermal gradient increasing linearly with depth from 10 to 60 °C. On the top and bottom surface corresponding conditions for hydraulic head h, salinity C and temperature T are held constant. In the simulation a heavy saltwater starts to enter on the top surface. It initializes cellular convective currents in the upper aquifer layer, where the saltwater sinks down, enters the abandoned borehole and salinates the lower aquifer layer. At the same time cooler water reaches the lower aquifer layer via the abandoned borehole. This thermohaline convection process is purely driven by the saltwater density and affected by thermal buoyancy.

The used model parameters and conditions are summarized in Table 14.5. We assume isotropic and homogeneous material conditions for each layer of the aquifer system. The flow in the abandoned borehole is described by Darcy law. The

[4] This test case was introduced by F. Cornaton in 2007 at the University of Neuchâtel, Center of Hydrogeology, Switzerland, with a number of unpublished simulations.

Table 14.5 Parameters and conditions used for the abandoned borehole problem

Quantity	Symbol	Value	Unit
Aquifer-aquitard-aquifer model domain shown in Fig. 14.31.			
Domain measure (width; depth; thickness)	$(L;\ H;\ B)$	$(100;\ 100;\ 100)$	m
Aquifer thicknesses	H_1	20	m
Aquitard thickness	H_2	60	m
Aquifer-aquitard-aquifer system			
Hydraulic conductivity (aquifers; aquitard)	$(K;\ K)$	$(1 \cdot 10^{-4};\ 1 \cdot 10^{-8})$	$\mathrm{m\,s^{-1}}$
Specific storage coefficient	S_o	$1 \cdot 10^{-6}$	$\mathrm{m^{-1}}$
Specific solutal expansion coefficient	α	0.2	1
Thermal expansion coefficient	β	$2 \cdot 10^{-4}$	$\mathrm{K^{-1}}$
Porosity (aquifers; aquitard)	$(\varepsilon;\ \varepsilon)$	$(0.2;\ 0.35)$	1
Molecular diffusion coefficient	D	$5 \cdot 10^{-10}$	$\mathrm{m^2\,s^{-1}}$
Volumetric heat capacity of (fluid; solid)	$(\rho c;\ \rho^s c^s)$	$(4.2 \cdot 10^6;\ 2.52 \cdot 10^6)$	$\mathrm{Jm^{-3}\,K^{-1}}$
Heat conductivity of (fluid; solid)	$(\Lambda;\ \Lambda^s)$	$(0.65;\ 3)$	$\mathrm{Jm^{-1}\,s^{-1}\,K^{-1}}$
Longitudinal dispersivity (mass and heat)	β_L	2	m
Transverse dispersivity (mass and heat)	β_T	0.2	m
Borehole DFE representation			
Flow law	Darcy (Table 4.5, case PN, 1D)		
Fracture area	A	1	$\mathrm{m^2}$
Isotropic hydraulic conductivity	K	0.1	$\mathrm{m\,s^{-1}}$
Specific storage coefficient	S_o	$1 \cdot 10^{-6}$	$\mathrm{m^{-1}}$
Specific solutal expansion coefficient	α	0.2	1
Thermal expansion coefficient	β	$2 \cdot 10^{-4}$	$\mathrm{K^{-1}}$
Porosity	ε	0.05	1
Molecular diffusion coefficient	D	$2.5 \cdot 10^{-13}$	$\mathrm{m^2\,s^{-1}}$
Volumetric heat capacity of (fluid; solid)	$(\rho c;\ \rho^s c^s)$	$(4.2 \cdot 10^6;\ 0)$	$\mathrm{Jm^{-3}\,K^{-1}}$
Heat conductivity of (fluid; solid)	$(\Lambda;\ \Lambda^s)$	$(0.65;\ 0)$	$\mathrm{Jm^{-1}\,s^{-1}\,K^{-1}}$
Longitudinal dispersivity (mass and heat)	β_L	0.1	m
Flow IC and BC's			
Initial condition (IC)	h_0	0	m
Dirichlet-type BC on top $(y = 0)$	h_{D_1}	0	m
Dirichlet-type BC at bottom $(y = -H)$[a]	h_{D_2}	-0.5	m
Salinity IC and BC's (normalized)			
Initial condition (IC) of salinity	C_0/C_s	0	1
Dirichlet-type BC on top $(y = 0)$	C_{D_1}/C_s	1	1
Dirichlet-type BC at bottom $(y = -H)$	C_{D_2}/C_s	0	1
Heat IC and BC's			
Reference temperature	T_0	10	°C
Initial condition (IC) of temperature	$f(y)$	$= -\frac{1}{2}y + T_0$	°C
Dirichlet-type BC on top $(y = 0)$	T_{D_1}	10	°C
Dirichlet-type BC at bottom $(y = -H)$	T_{D_2}	60	°C
FEM			
Uniform 2D and 3D meshes of $(10^2 \cdot 2^\ell)^D$ linear quadrilateral and brick elements, respectively, containing $60 \cdot 2^\ell$ 1D DFE's $(\ell = 0, 1, 2;\ D = 2, 3)$, GFEM (no upwind), OB approximation			

(continued)

Table 14.5 (continued)

Quantity	Symbol	Value	Unit
Initial time step size[b]	Δt_0	10^{-5}	d
RMS error tolerance (AB/TR)	ϵ	10^{-4}	1
Simulation time period	t_{end}	365	d

[a] Hydrostatic condition is assumed at bottom by using (L.16) of Appendix L with $h_0 = C_{D_2} = 0$, $T_s = 60$ °C, $T_0 = 10$ °C

[b] In addition, maximum rate of time step change $\Xi = \frac{\Delta t_{n+1}}{\Delta t_n} = 2$ and maximum time step size $\Delta t_{\max} = 0.5\,\text{d}$

simulations covering a time of 1 year are fully transient both for flow, saltwater and heat transport. BC's unreported in Table 14.5 for flow, saltwater and heat transport represent boundaries at which natural BC's are imposed, i.e., $-(\boldsymbol{K} \cdot \nabla h) \cdot \boldsymbol{n} = 0$, $-(\boldsymbol{D} \cdot \nabla C) \cdot \boldsymbol{n} = 0$ and $-(\boldsymbol{\Lambda} \cdot \nabla T) \cdot \boldsymbol{n} = 0$, respectively.

We simulate the thermohaline convection process by using both 2D and 3D models with different spatial resolutions. Regular meshes of linear quadrilateral elements in 2D and linear brick elements in 3D are chosen. With increasing mesh refinement level of $\ell = 0, 1, 2, \ldots$, the resulting number of elements N_{E} and nodes N_{P} are

$$\begin{aligned} N_{\text{E}} &= (10^2 \cdot 2^\ell)^D \\ N_{\text{P}} &= (10^2 \cdot 2^\ell + 1)^D \end{aligned} \tag{14.45}$$

where $D = 2, 3$ represents the dimension. The abandoned well is embodied in the meshes by using $60 \cdot 2^\ell$ 1D DFE's both in 2D and 3D schematizations. For all FEFLOW simulations we use GFEM (without any upwinding), adaptive GLS 2nd-order accurate predictor-corrector AB/TR time integrator, FKA consistent velocity and OB approximation. Comparisons will be given to the computational results obtained by the finite-element research code Ground Water (GW) [102] using same mesh and DFE data.

Due to the layer structure and the presence of hydrodynamic dispersion it is obvious that the quantification of the convective regime via solute and thermal Rayleigh numbers, Ra_c (11.25), Ra_t (11.26), is not possible. In particular, the dispersivities β_L, β_T introduce additional nonlinear dependences of saltwater mixing and thermal conduction on the convective velocity. If we disregard dispersivity effects (acceptable at initial phase) and consider only the top aquifer layer we can make a rough estimate from a HRL problem equivalence (cf. Sect. 11.5) and assess a solutal Rayleigh number of $\text{Ra}_c = -4 \cdot 10^6$ and a thermal Rayleigh number of $\text{Ra}_t = 33$, which clearly indicate a monotonic convection representing a fingering regime in the CSA quadrant of the DDC stability diagram of Fig. 11.8. A Turner number (11.29) of $\text{Tu} = 100$ indicates the gravitational dominance of the saltwater. As a consequence, we must expect a strong primarily solute-driven free convection behavior which is sensitive to inherent perturbation and discretization effects. We note, however, that dispersion effects can significantly reduce $|\text{Ra}_c|$

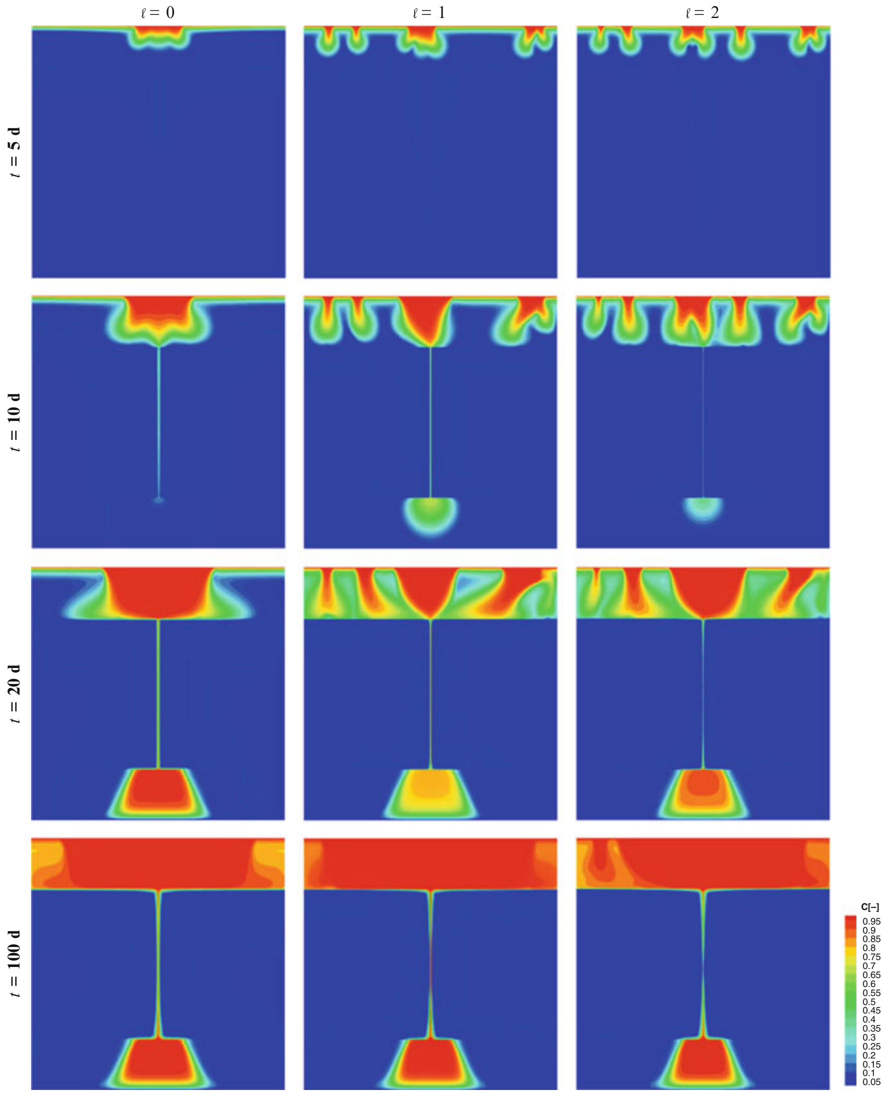


Fig. 14.32 Salinity patterns for 2D meshes of refinement levels $\ell = 0, 1, 2$ ($N_{\mathrm{E}} = 10^4$, $4 \cdot 10^4$, $1.6 \cdot 10^5$) and different times $t = 5, 10, 20, 100$ (d) simulated by FEFLOW. Color sequence *blue-lightblue-green-yellow-orange-red* depicts normalized salinity C/C_s from 0 to 1 using 20 intervals

and $|\mathrm{Ra}_t|$ because the effective 'diffusion' increases with $\varepsilon D + \beta_L |q_c|$, where the density-dependent Darcy velocity $|q_c|$ could be in the range $0 \leq |q_c| \lesssim K$.

The evolution of salinity and temperature for 2D meshes of three consecutive refinement levels $\ell = 0, 1, 2$ simulated by FEFLOW is shown in Figs. 14.32

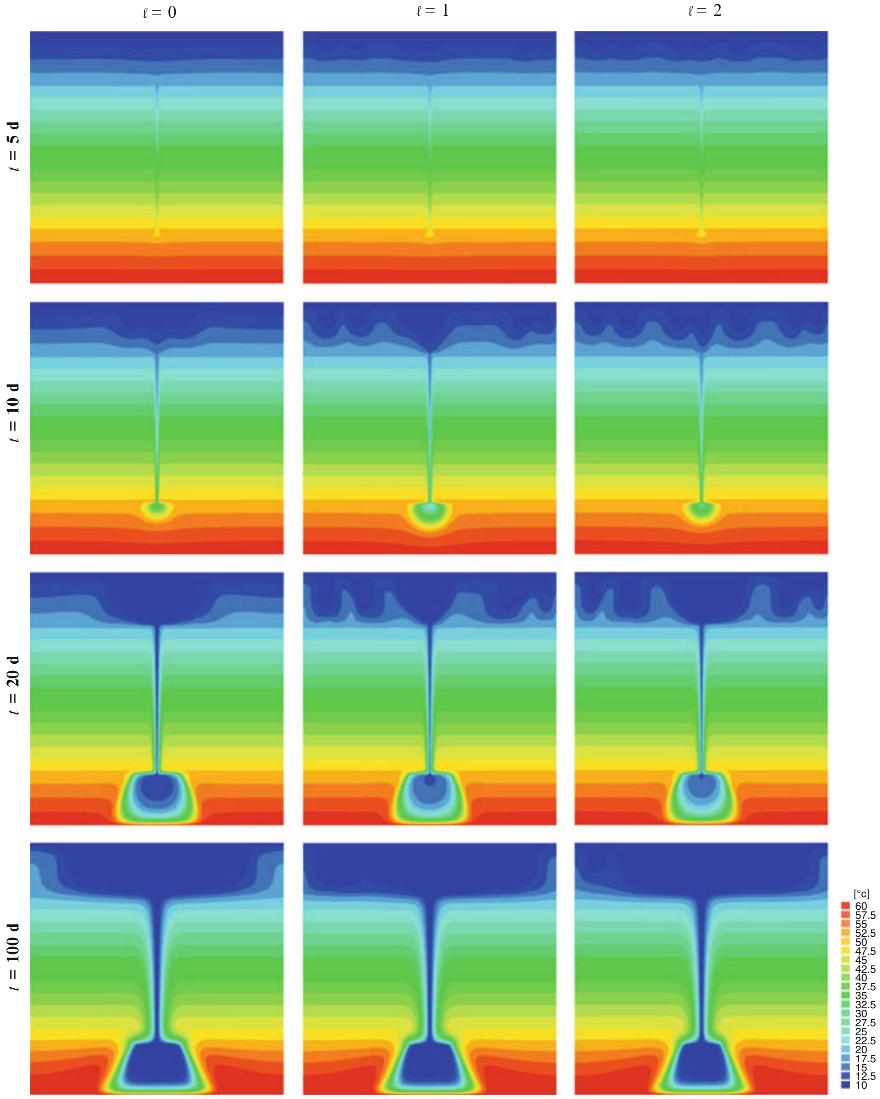


Fig. 14.33 Temperature patterns for 2D meshes of refinement levels $\ell = 0, 1, 2$ ($N_E = 10^4, 4 \cdot 10^4, 1.6 \cdot 10^5$) and different times $t = 5, 10, 20, 100$ (d) simulated by FEFLOW. Color sequence *blue-lightblue-green-yellow-orange-red* depicts temperature T from 10 to 60 °C using 20 intervals

and 14.33, respectively. It clearly reveals the dependence of the spatial resolution on the convection process. While a coarser mesh with $\ell = 0$ produces symmetric patterns, more refined meshes lead always to unsymmetric patterns in the salinity and, correspondingly, in the temperature field. It is obvious, a higher resolution

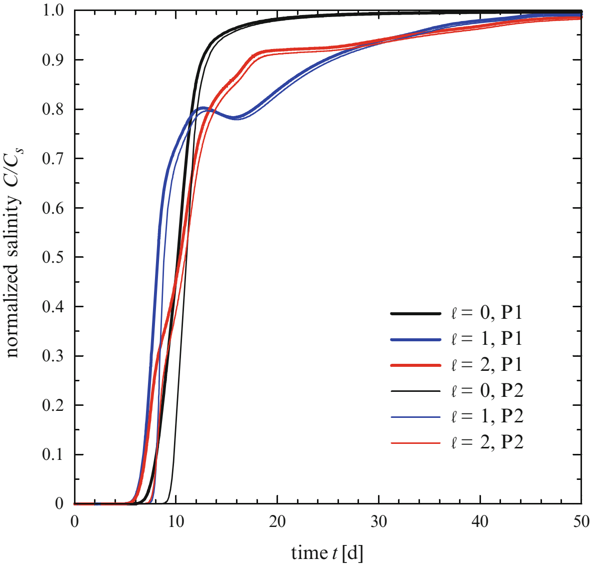


Fig. 14.34 Salinity breakthrough curves at entry point point P1$(x, y) = (50\,\text{m}, -20\,\text{m})$ and exit point P2$(x, y) = (50\,\text{m}, -80\,\text{m})$ of the abandoned borehole simulated by FEFLOW for 2D meshes of refinement levels $\ell = 0, 1, 2$

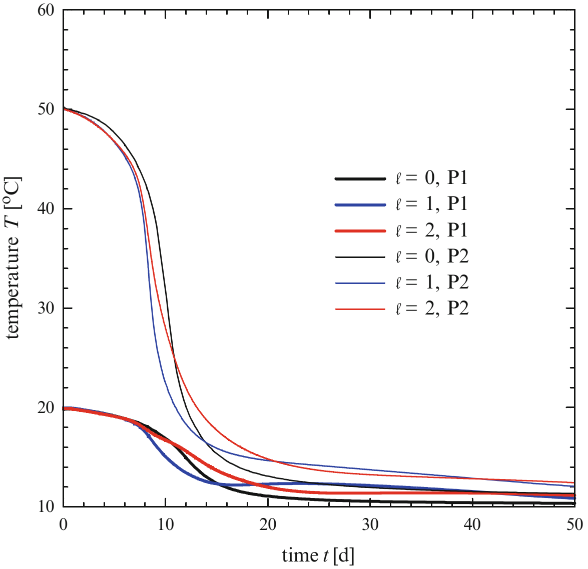


Fig. 14.35 Temperature breakthrough curves at entry point point P1$(x, y) = (50\,\text{m}, -20\,\text{m})$ and exit point P2$(x, y) = (50\,\text{m}, -80\,\text{m})$ of the abandoned borehole simulated by FEFLOW for 2D meshes of refinement levels $\ell = 0, 1, 2$

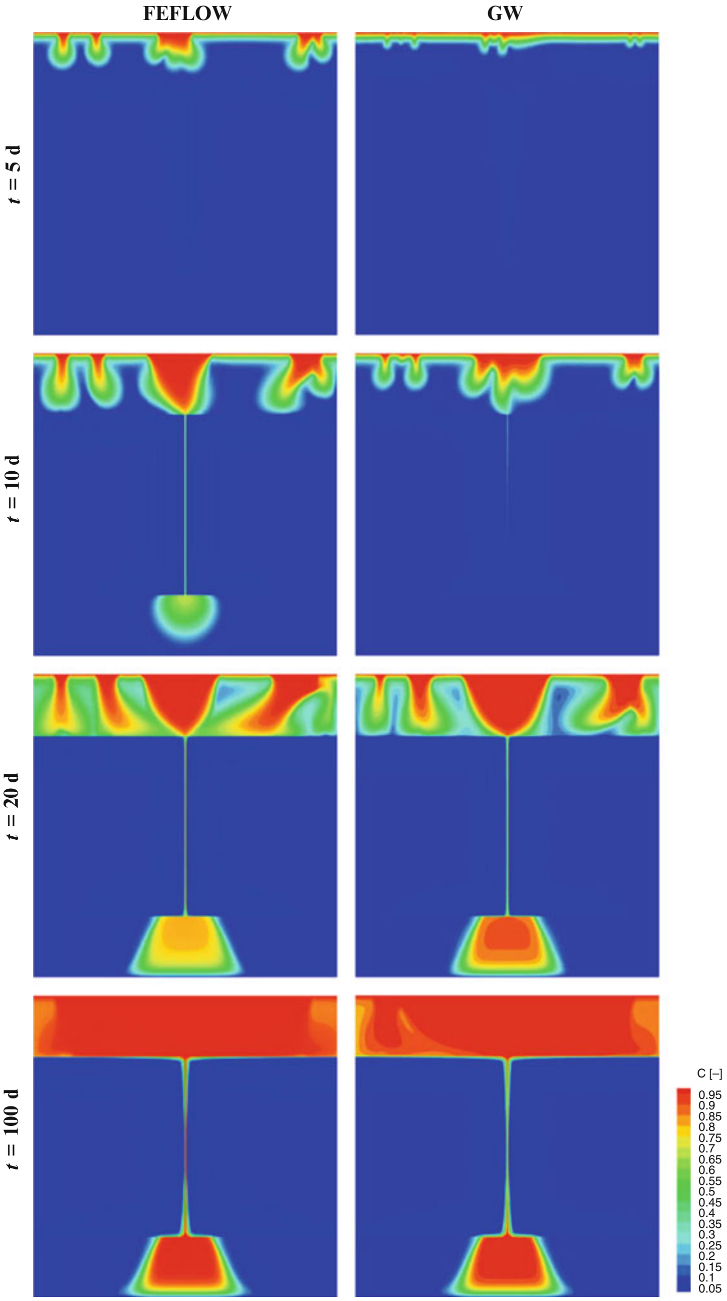


Fig. 14.36 Comparison between FEFLOW's and GW's salinity patterns for 2D mesh of refinement level $\ell = 1$ ($N_{\mathrm{E}} = 4 \cdot 10^4$) and different times $t = 5, 10, 20, 100$ (d). Color sequence *blue-lightblue-green-yellow-orange-red* depicts normalized salinity C/C_s from 0 to 1 using 20 intervals

Fig. 14.37 Comparison between FEFLOW's and GW's salinity breakthrough curves at entry point point P1$(x, y) = (50\,\text{m}, -20\,\text{m})$ of the abandoned borehole for 2D meshes of refinement levels $\ell = 0, 1$

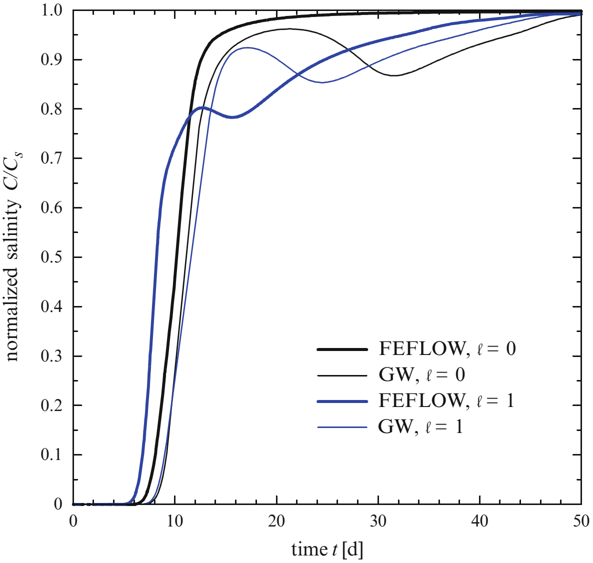
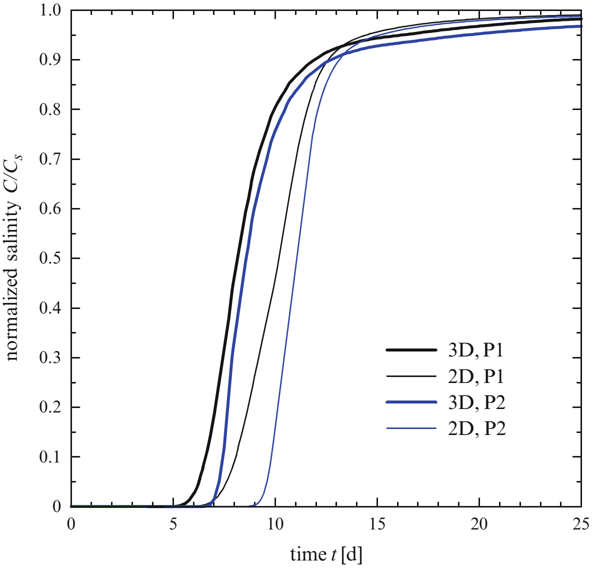


Fig. 14.38 Comparison of salinity breakthrough curves at entry point point P1$(x, y, z) = (50\,\text{m}, -20\,\text{m}, 0\,\text{m})$ and exit point P2$(x, y, z) = (50\,\text{m}, -80\,\text{m}, 0\,\text{m})$ of the abandoned borehole simulated by FEFLOW for 3D and 2D meshes of refinement level $\ell = 0$



implies more inherent perturbing noise, which triggers the convective instability in the upper boundary layer of salinity at certain locations in a random manner. Notice, for the present simulations we do not induce extra perturbations on the top boundary. As illustrated in Fig. 14.32 the salinity reaches the bottom of the upper aquifer layer after about 10 days and leads to a breakthrough of salinity in the abandoned borehole. Once saltwater enters the borehole a fast descent into the lower

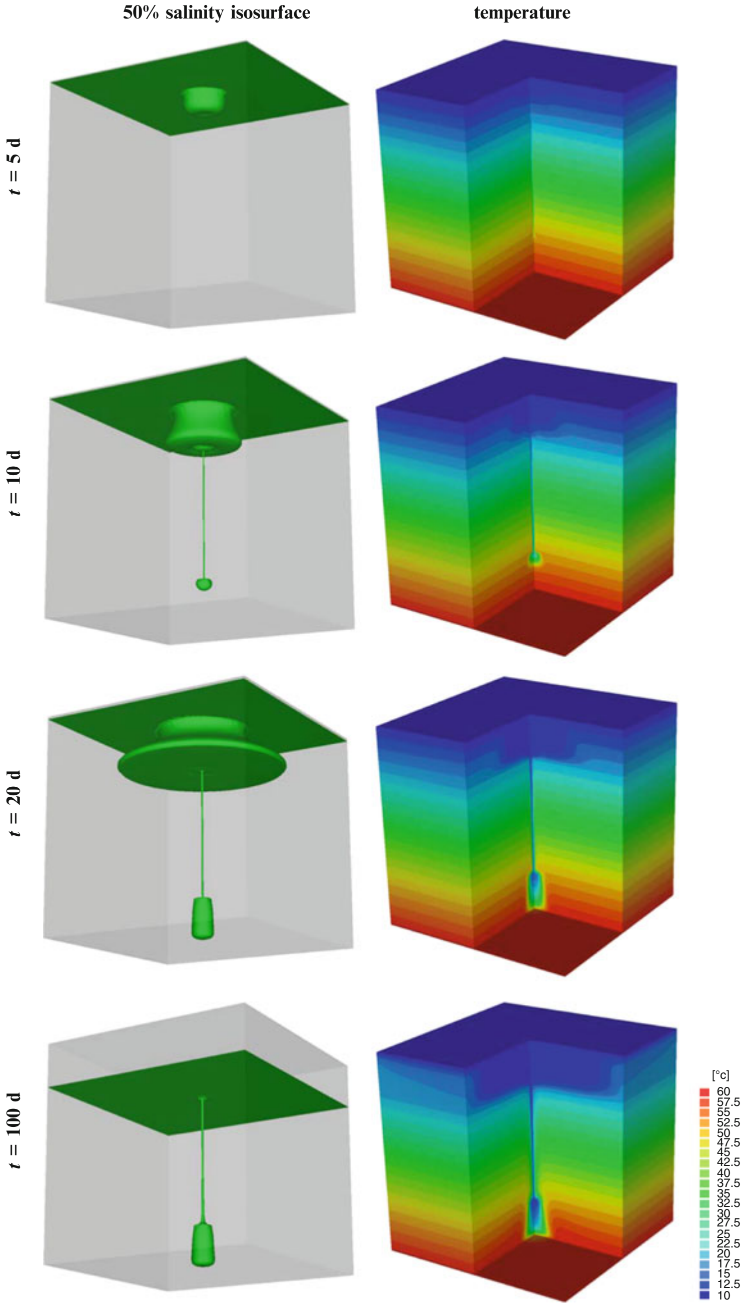


Fig. 14.39 Fifty percentage salinity isosurface and temperature field for 3D mesh of refinement level $\ell = 0$ ($N_E = 10^6$) at different times $t = 5, 10, 20, 100$ (d) simulated by FEFLOW. Color sequence *blue-lightblue-green-yellow-orange-red* depicts temperature T from 10 to 60 °C using 20 intervals

aquifer occurs, where saltwater spreads conically over time. On the other hand, the temperature field features a negative image to the salinity pattern (Fig. 14.33). With the sinking of heavy saltwater the aquifer layers and the borehole are cooled down. It is remarkable that fingering convection only occurs at beginning in the upper aquifer layer. At later times this effect vanishes and the solution approaches to a steady state equivalent for all mesh resolutions. As a consequence, the saltwater and temperature breakthrough in the borehole at beginning is determined by the history of convection in the upper aquifer layer, which implies mesh dependency as evidenced in Figs. 14.34 and 14.35. In dependence on the actual history of free convection developing in the upper aquifer the simulated breakthrough curves can be nonmonotonic and lagged.

In the FEFLOW simulations the number of AB/TR adaptive time steps took about 1,200 for refinement level $\ell = 0$ (both in 2D and 3D), about 1,800 for for refinement level $\ell = 1$ and about 3,700 for for refinement level $\ell = 2$. In Fig. 14.36 FEFLOW's salinity results for the 2D mesh with refinement level $\ell = 1$ are compared to the findings obtained by the finite element simulator GW [102]. It indicates that both codes simulate quite different convection patterns at beginning. It is obvious that in FEFLOW's computations the finger evolution is faster and the resulting saltwater breakthrough in the borehole is more advanced. This is also shown in the breakthrough curves of Fig. 14.37 for the two refinement levels $\ell = 0, 1$.

FEFLOW simulations are also performed for the equivalent 3D problem by using a mesh of refinement level $\ell = 0$ ($N_E = 10^6$). The 3D breakthrough histories in comparison to 2D are given in Fig. 14.38 for $\ell = 0$. It reveals that the breakthrough in 3D is clearly faster than in 2D. The developments of salinity and temperature for the 3D model are shown in Fig. 14.39. It illustrates how the heavier and cooler saltwater intrudes very locally via the tubular borehole.

Chapter 15
Specific Topics

15.1 Finite Element Meshing

15.1.1 General

The finite element solution of the governing flow, mass and heat transport equations as described in the preceding chapters requires the discretization of the equations to replace the continuous PDE's with a system of simultaneous algebraic equations (cf. Chap. 8). The spatial discretization is accomplished by subdividing the study domain with its boundary into a number of nonoverlapping finite elements of different shapes, such as triangles, tetrahedra, bricks (see Fig. 8.6), forming the *finite element mesh* associated with a set of *nodes* and interpolation functions. Such a mesh should be sufficiently dense and appropriately refined according to the changes in the solution gradients to obtain accurate numerical approximations. On the other hand, the constructed mesh should ensure computational efficiency and robustness, e.g., in avoiding too distorted element shapes and discontinuous changes.

Meshes can be classified according to

(i) CONFORMITY. We can differ between *conformal* and *nonconformal* meshes (Fig. 15.1). Conformal meshes are characterized by a perfect match of edges and faces between neighboring elements, while nonconforming meshes do not match perfectly between neighboring elements and give rise to so-called *hanging nodes*. A conformal discretization can be attained on a nonconforming mesh by using constraints and associated interpolation functions. However, such type of interpolation must not satisfy local conservativity at those hanging nodes. Although nonconforming meshes are very convenient in handling adaptive local mesh refinements, their use can reduce accuracy and stability. In the present work we generally prefer conformal meshes.

(ii) ALIGNMENT. We differ between surface alignment of meshes whose boundary faces match perfectly the surface (or interfaces) of the domain and nonsurface aligned meshes in which faces are crossed by the surface (Fig. 15.2).

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5_15,

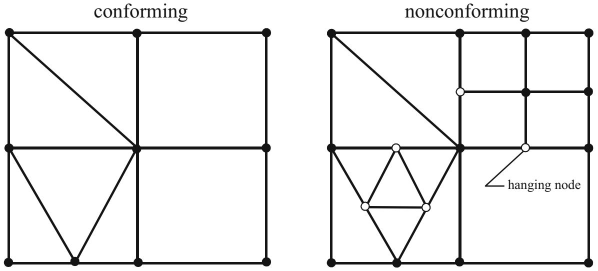


Fig. 15.1 Conformal versus nonconformal mesh

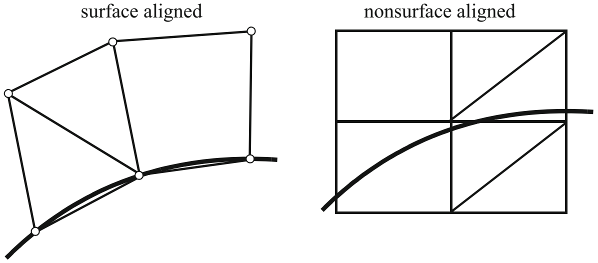


Fig. 15.2 Surface aligned versus nonsurface aligned mesh

An important advantage of FEM is in working with surface alignment, where outer and inner boundaries or material interfaces and zones can be correctly meshed.

(iii) TOPOLOGY. We differ between *structured* and *unstructured* meshes (Fig. 15.3). A mesh is called structured if each nodal point has the same number of neighbors (except for nodes on boundaries) so that the nodes can be ordered into a regular index array (i, j, l) with the assumption that the nodes (i, j, l), $(i + 1, j, l)$, etc., are neighbors, representing for instance a FDM-like stencil of quadrilateral or hexahedral configurations. Contrarily, an unstructured mesh possesses nodes which can have an arbitrary number of neighbors, typically in using triangular or tetrahedral elements. There are also composite meshes termed as multiblock meshes, where the mesh is assembled from groups of structured submeshes of quadrilaterals or bricks, forming together an unstructured mesh. FEM is superior in using unstructured meshes providing a maximum in geometric flexibility.

The finite element meshing process starts in preparation of all important geometric entities and input data describing the study domain with its boundaries, material zones, subregions, interfaces and local points on which specific conditions and parameters must be assigned. This work can be very comprehensive and time-critical. There are in principle two ways of describing the required geometric entities possibly linked to input data:

- *Using analytical functions.* This technique is practiced in engineering industry by using CAD/CAM systems where the problem is usually described by a composition of analytical, semianalytical and/or parametric-based entities. Splines, B-splines, NURBS or other types of functions can be used to define the surface

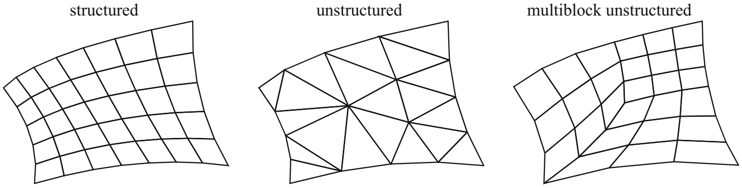


Fig. 15.3 Different mesh topology

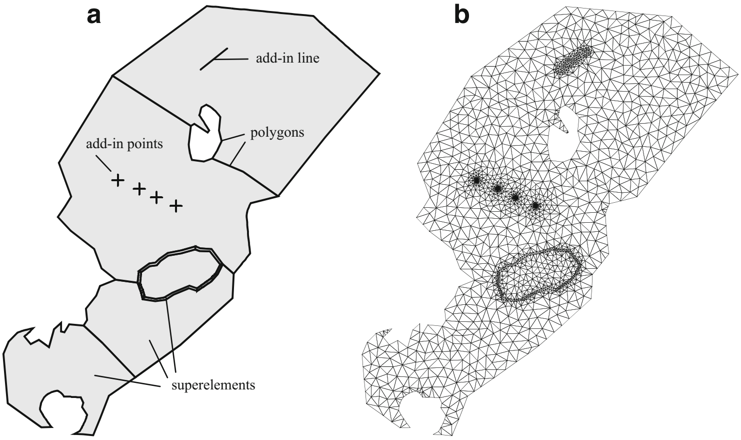


Fig. 15.4 (**a**) Input superelement mesh: boundaries, material zones and BC locations of a 2D computational domain are described by a number polygons, add-in lines and points in form of superelements. (**b**) Unstructured finite element mesh resulting from a suitable triangulation of the superelements

of the domain providing an explicit and continuous representation, which can form a direct input for the finite element meshing process [514].

- *Using discrete data.* This technique is usually preferred in geosciences by using GIS systems [155, 419, 434] and 3D geologic modeling [263], where an analytical and continuous representation of the geometric data is usually not given. GIS handles digital data in (1) vector or (2) raster form, which are geographically referenced as maps. Vector data are the set of points, lines and polygons that are used to represent map feature locations. Raster data are described as a grid of square or rectangular data. GIS allows the storage, manipulation, analysis and visualization of a large volume of data. It provides database functionality and can easily maintain and update spatial data with their associated information. An important feature of GIS in the present context is that database information linked to the geometric entities can be directly exploited in the meshing process for assigning material data, BC's and other quantities to the corresponding locations in the discretized model domain. A 2D example is shown in Fig. 15.4, where polygon, line and point data are collected to so-called *superelements*, which form the input of the mesh generation. The assembly of all superelements is termed as *superelement mesh*.

15.1.2 Mesh Generation

A number of powerful methods of generating meshes that are suited to particular applications have been developed [188, 514]. We shall describe those strategies which are most useful for the present class of porous-media and fracture modeling, available in FEFLOW.

15.1.2.1 Transport Mapping

Transport mapping represents a relatively simple, but fast and robust meshing method to generate structured or multiblock unstructured finite element meshes [165, 188, 314, 514]. It is commonly restricted to domains which are of quadrilateral or hexahedral shape or can be composed of a number of quadrilaterals or hexahedra (e.g., cross section, rectangular domain, regularly layered structure).

Let us briefly describe the method in 2D. A global quadrilateral domain or a superelement of quadrilateral shape is described by four sides. In doing so, we can transform the physical domain given in Cartesian coordinates $\boldsymbol{x}$ to a much simpler square domain given in local coordinates $\boldsymbol{\eta}$. This coordinate transformation provides a one-to-one mapping between the physical $\boldsymbol{x}-$space and the computational $\boldsymbol{\eta}-$space, viz.,

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\eta}) \tag{15.1}$$

provided that the Jacobian $|\boldsymbol{J}| = |\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\eta}}|$ is nonzero.

To treat curved boundaries we chose curvilinear geometries and prefer a biquadratic isoparametric map which is identical to the transformation used for the curved 8-node biquadratic finite element (cf. Table G.2c of Appendix G). As shown in Fig. 15.5 such type of mapping allows the construction of a biparabolic geometry in the physical coordinates $\boldsymbol{x}$ with prescribed local coordinates $\boldsymbol{\eta}$ similar to (8.71)

$$\begin{aligned} x &= \textstyle\sum_{J=1}^{8} N_J(\xi,\eta) x_J \\ y &= \textstyle\sum_{J=1}^{8} N_J(\xi,\eta) y_J \end{aligned} \tag{15.2}$$

where (x_J, y_J) are the 2D Cartesian coordinates of the 8 nodes describing the curved sides of the quadrilateral superelement and N_J are the isoparametric shape functions listed in Table G.2c of Appendix G.

In practice, the mesh is generated with less computational effort by subdividing regularly the computational domain along the $\xi-$ and $\eta-$directions with given increments $\Delta\xi$ and $\Delta\eta$, respectively, according to a chosen number of rows NR > 1 and columns NC > 1 for the quadrilateral sides

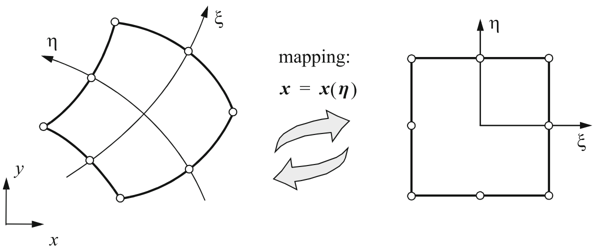


Fig. 15.5 Mapping of a quadrilateral superelement with parabolic sides

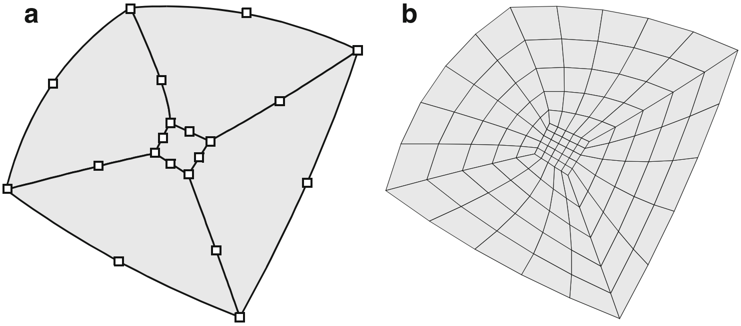


Fig. 15.6 (**a**) Superelement mesh consisting of five quadrilaterals, where their parabolic sides are specified by corner and midside (superelement) nodes. (**b**) Resulting all-quadrilateral element mesh by using transport mapping for a 5 × 5 subdivision of each quadrilateral superelement

$$\Delta\xi = \frac{2}{\mathrm{NR}-1}, \quad \Delta\eta = \frac{2}{\mathrm{NC}-1} \tag{15.3}$$

where $(-1 \leq \xi \leq 1)$ and $(-1 \leq \eta \leq 1)$. It leads to an array of NR × NC local coordinates (ξ_i, η_j) with $\xi_i = -1 + (i-1)\Delta\xi$, $1 \leq i \leq \mathrm{NR}$ and $\eta_j = -1 + (j-1)\Delta\eta$, $1 \leq j \leq \mathrm{NC}$. This regular array of local coordinates in the $\boldsymbol{\eta}-$space is input in the coordinate transformation (15.2) to compute the coordinates in the physical space $\boldsymbol{x}$:

$$\left.\begin{aligned} x_{(i,j)} &= \textstyle\sum_{J=1}^{8} N_J(\xi_i, \eta_j) x_J \\ y_{(i,j)} &= \textstyle\sum_{J=1}^{8} N_J(\xi_i, \eta_j) y_J \end{aligned}\right\} \quad (1 \leq i \leq \mathrm{NR},\ 1 \leq j \leq \mathrm{NC}) \tag{15.4}$$

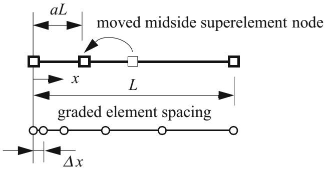
For each quadrilateral superelement the mapping (15.4) produces a regular mesh of (NR − 1) × (NC − 1) quadrilateral elements. The procedure is performed for each quadrilateral of a given superelement mesh such as exemplified in Fig. 15.6 for five quadrilateral superelements, each subdivided simply by 5 × 5 quadrilaterals.

This type of biquadratic mapping also allows stretching and compression of points along the sides of the superelement quadrilaterals so that specific regions of the domain can be resolved more accurately. It can be done easily by moving the midside superelement node along the side up to a maximum displacement of the quarter of the side length such as shown in the example of Fig. 15.7. The midside node movement should not equal to or larger than the quarter of the side length, otherwise the transformation Jacobian could be no more positive and the mapping must fail[1].

There are cases where the superelement sides cannot suitably be transformed via a simple biquadratic isoparametric map (15.2), for example one or two sides form circular arc segments are given in polar (cylindrical) coordinates (Fig. 15.8). The easiest extension is the use of algebraic mapping [165, 314, 514] in which boundary data are interpolated to generate the interior mesh that is orthogonal or near-orthogonal adjacent to the bounding lines (surfaces). The standard method is known as *transfinite interpolation* [165, 314]. In 2D the four sides $\widehat{12}$, $\widehat{23}$, $\widehat{34}$ and $\widehat{41}$ of a quadrilateral superelement are described by suited parametric equations $F_{12}(\boldsymbol{\eta})$, $F_{23}(\boldsymbol{\eta})$, $F_{34}(\boldsymbol{\eta})$ and $F_{41}(\boldsymbol{\eta})$, respectively, as function of the parametric (local) coordinates $\boldsymbol{\eta}$. Internal mesh lines are constructed by linear interpolation between opposite bounding lines followed by an orthogonal adjustment of the generated points. The linear interpolation can be combined by stretching function $s(\xi)$ such as [165]

$$s = P\xi + (1-P)\left(1 - \frac{\tanh\left[Q(1-\xi)\right]}{\tanh Q}\right) \tag{15.5}$$

[1] Considering a superelement side of length L and move the midside node to the distance aL, where $0 \leq a \leq \frac{1}{2}$ is a shifting factor, viz.,



the smallest element length Δx of the graded element spacing obtained with the parabolic mapping is:

$$\Delta x = \Delta\xi L[\Delta\xi(\tfrac{1}{2} - a) + 2a - \tfrac{1}{2}]$$

where $\Delta\xi > 0$ is the given increment (15.3) of the superelement side subdivision. It results by taking the parabolic interpolation functions of Tab. G.1(b) of Appendix G with $\xi = -1 + \Delta\xi$ for the second evaluation point. For $a = \frac{1}{2}$ the standard equally graded spacing with $\Delta x = \frac{1}{2}\Delta\xi L$ is given, while with the a midside node shift of $a = \frac{1}{4}$ a left-sided densification with $\Delta x = \frac{1}{4}\Delta\xi^2 L$ results. Since $\Delta\xi(\frac{1}{2} - a) + 2a - \frac{1}{2}$ must be positive, the following constraints are required:

$$\begin{array}{rcll} a & > & \frac{1}{2}\left(\frac{1-\Delta\xi}{2-\Delta\xi}\right) & \text{for} \quad 0 \leq \Delta\xi \leq 1 \\ \Delta\xi & > & \frac{1-4a}{1-2a} & \text{for} \quad 0 \leq a \leq \frac{1}{4} \end{array}$$

We recognize that with decreasing $\Delta\xi \to 0$ the shift of the midside node must satisfy $a > \frac{1}{4}$.

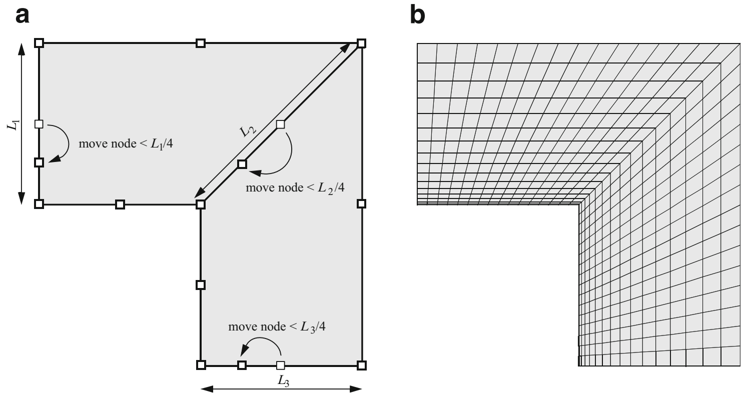


Fig. 15.7 (**a**) Midside nodes along the sides of the superelement are moved up to the quarter of the side lengths L_x to densify meshing. (**b**) Resulting mesh with compressed elements at boundary for a movement of $L_x/4.1$

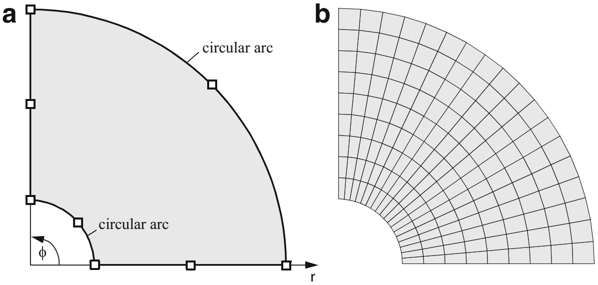


Fig. 15.8 (**a**) Annular superelement bounded by two circular arc segments. (**b**) Resulting all-quadrilateral element mesh by using transfinite interpolation for a 9 × 18 subdivision

where $0 \leq \xi \leq 1$ is a normalized coordinate, P and Q are free parameters to provide mesh point control. In FEFLOW's implementation P is set in dependence on the shift of the middle node along a superelement side and is taken in the range of 10^{-3} and 1.9, while Q is set to 2. An example of mesh generation by using transfinite interpolation is shown in Fig. 15.8b for a quadrilateral superelement which consists of two bounding circular arcs.

Transport mapping has proved to be a fast and robust method to generate regular, structured or multiblock unstructured, all-quadrilateral and all-hexahedral element meshes in 2D and 3D, respectively. This geometric regularity can be desired in various applications. Sometimes, however, a regular triangular mesh is more favorable to attain a higher geometric flexibility for a subsequent local refinement. For this need it is possible to partition a quadrilateral mesh by triangles. Each quadrilateral element can be split into two triangles with respect to the shortest diagonal or by specifying the diagonal or can be split into four triangles such as exemplified in Fig. 15.9.

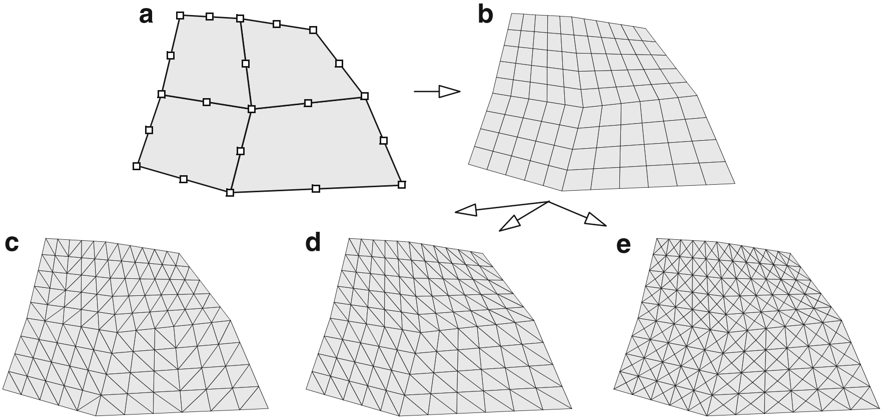


Fig. 15.9 (**a**) Basic quadrilateral superelement mesh, (**b**) resulting all-quadrilateral element mesh, (**c**) two-triangle splitting with respect to the shortest diagonal, (**d**) two-triangle splitting with respect to a specified diagonal and (**e**) four-triangle splitting

15.1.2.2 Advancing Front Technique

Advancing front technique (AFT) is very attractive to generate unstructured triangular meshes in 2D and tetrahedral meshes in 3D [188,353,514]. It constructs the mesh of the domain from its boundary and can handle rather complex geometric shapes. While AFT offers a great geometric flexibility in the mesh generating process, the generated number of points and elements vary through the domain in dependence on the boundary geometry, the number of used boundary points and their distributions. As a consequence, a lack of regularity results in the mesh and, contrary to the transport-mapping method (Sect. 15.1.2.1), the number of elements of the final mesh is not assessable in advance.

The AFT generation process starts by discretizing each boundary curve. Typically in 2D, at the start the boundary consists of the sequence of straight line segments of a superelement that connect consecutive boundary nodes.[2] It represents the initial *front*. This generation front is used to create internal triangles in

[2] Boundary nodes $\boldsymbol{x}_i$ ($i = 1, 2, \ldots$) are created on each side of a superelement. Their distances depend on the desired element resolution. They can be equally distributed along the superelement side or can be densified locally by using a parabolic grading function:

$$\boldsymbol{x}_i = \boldsymbol{x}_2 + \xi_i(k)[\boldsymbol{a} + \xi_i(k)\boldsymbol{b}], \quad (-5 \le k \le 5)$$

with

$$\boldsymbol{a} = \tfrac{1}{2}(\boldsymbol{x}_3 - \boldsymbol{x}_1), \quad \boldsymbol{b} = \tfrac{1}{2}(\boldsymbol{x}_3 + \boldsymbol{x}_1) - \boldsymbol{x}_2$$

$$\xi_i(k) = \eta_i - s(k)(\eta_i^2 - 1), \quad \eta_i = -1 + (i-1)\Delta\eta, \quad s(k) = \tfrac{1}{4}\mathrm{sgn}(k)\sum_{j=1}^{|k|} 2^{-j+1}$$

dependence on the geometric properties of the segments. The length of the segments must be consistent with the desired local distribution of mesh size. Let α be the angle formed by two consecutive segments of the front, then three major patterns for concave and convex geometry can be identified [188]:

- Pattern (a): $\alpha < \frac{\pi}{2}$, the two segments with the angle α form two edges of a single triangle created (Fig. 15.10a).
- Pattern (b): $\frac{\pi}{2} \leq \alpha \leq \frac{2\pi}{3}$, an internal point is created and two triangles are generated by using the two segments with angles α (Fig. 15.10b).
- Pattern (c): $\frac{2\pi}{3} < \alpha$, one segment is retained, a triangle is created with this segment as an edge and an internal point (Fig. 15.10c).

The position of the internal points are chosen to obtain new triangles as equilateral as possible. At each point creation it must be verified that the point is inside the domain, but not inside an existing element. This verification is most crucial for AFT, in particular for 3D where the front consists of triangular faces [188, 353, 514].

While the creation of internal triangles progresses, the generation front is updated in such a way that only segments remain part of the front which are available to form an element side of further internal triangles in a next step. Accordingly, the generation front changes continuously and needs to be updated whenever a new element is created. The AFT generation process finishes when the front becomes empty, i.e., when the domain is completely filled by triangles. An example of AFT meshing is illustrated in Fig. 15.11 for a circular 2D domain showing the initial front and the form of the mesh at various stages during the AFT generation process. We recognize that AFT can advance geometric irregularity in the mesh starting from the initial front. Note also that AFT may fail for pathological cases, for example if the point-discretized starting front is chosen too coarse or too sharply varied so that it becomes inconsistent with a complexly shaped boundary. To get better control of the interior meshing behavior and overcome possible weaknesses it is useful to subdivide the computational domain into a larger number of simpler-shaped superelements, in particular if the domain has a complicate geometry. For each superelement of a superelement mesh the AFT triangulation is performed and even very complex domains can be successfully and quickly meshed.

There are many variants of AFT [188, 514]. For example, the automatic mesh generator GRIDBUILDER [369] can belong to this category in a broader sense, which is loosely based on Sadek's approach [455] using a specific technique for advancing triangulation from boundaries of the subdivided domain, however, restricted only to 2D geometries.

where $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are the coordinates of the left, middle and right nodes of a superelement side, respectively, k is a grading counter (for $k = 0$ there is no grading and the nodes become equally distributed, $k > 0$ leads to left-sided densification, $k < 0$ leads to a right-sided densification of boundary nodes) and $\Delta\eta = 2/(\mathrm{NS} + 1)$ is a local coordinate increment determined by the desired number of superelement side segmentation NS.

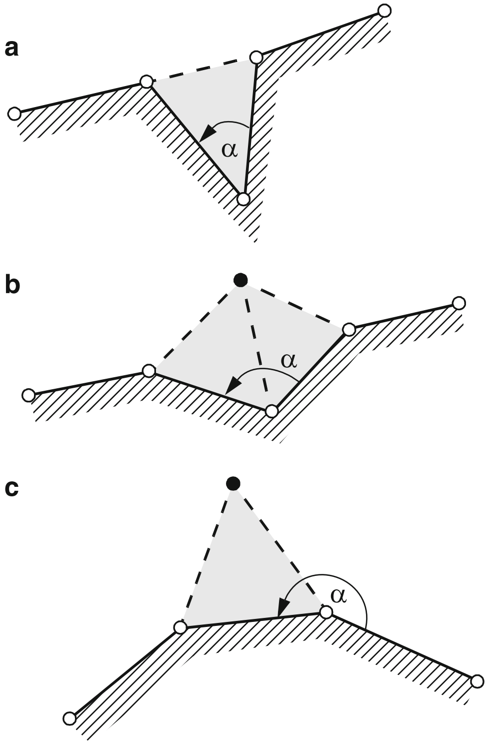


Fig. 15.10 AFT construction of triangles and internal points from a 2D generation front for the three major patterns (**a**)–(**c**) (Modified from [188])

15.1.2.3 Delaunay-Voronoï Method

The *Delaunay-Voronoï method* (DVM) represents a very powerful and fertile meshing concept to generate unstructured triangular or tetrahedral meshes [188,353,514]. Given a set of points $\mathcal{P} := \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, we may subdivide the space into regions or volumes $\mathcal{V} := V_1, V_2, \ldots, V_n$ assigned to each of the points in such a way that any location $\boldsymbol{x}$ within V_i is closer to $\boldsymbol{x}_i$ than to any other of the points, viz.,

$$V_i = \Big\{\mathcal{P} : \|\boldsymbol{x} - \boldsymbol{x}_i\| < \|\boldsymbol{x} - \boldsymbol{x}_j\|\Big\}, \quad \forall j \neq i \tag{15.6}$$

This tessellation $\mathcal{V}$, which covers the domain completely, results in a set of nonoverlapping convex regions called *Voronoï regions* forming convex polygons in 2D and convex polyhedra in 3D. The sum of all points $\boldsymbol{x}$ satisfying (15.6) defines such a Voronoï region.

In 2D it is easy to recognize that a bounding side of a Voronoï polygon must be midway between the two points $\boldsymbol{x}_i, \boldsymbol{x}_j$ and form thus a segment of the perpendicular bisector of the line joining these two points. If all the pairs of points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are joined by straight lines, a triangulation of the convex hull of $\mathcal{P}$ results. This

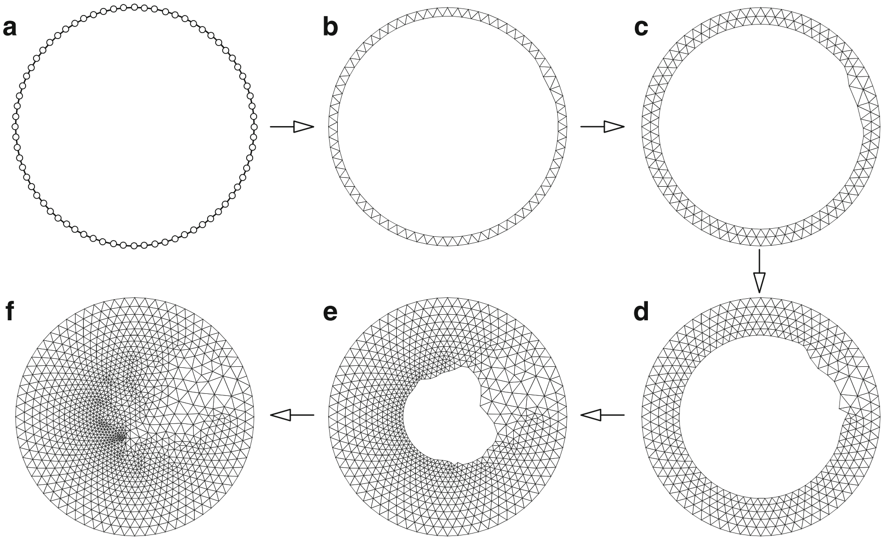


Fig. 15.11 AFT generation showing different stages during the triangulation process for a circular 2D domain: (**a**) discretized outer boundary forming the initial front, (**b**)-(**e**) intermediate fronts advancing into the circular domain, (**f**) final unsmoothed mesh

triangulation is known as the *Delaunay triangulation*. An example of a 2D Delaunay triangulation is shown in Fig. 15.12. Equivalent geometric construction exists in 3D where a set of tetrahedra results from joining the points across polyhedral boundaries of Voronoï volumes [188, 353, 514].

An important property of any Delaunay triangulation is the in-circle criterion. It states that no other point is contained within the circumcircle (circumsphere) formed by the nodes of the triangle (tetrahedron), valid in arbitrary dimensions. This property is used to construct algorithms for the triangulation. By satisfying the in-circle criterion a Delaunay triangulation possesses further attractive properties such as ensuring angle regularity of the triangles (tetrahedra) which has positive consequences in the discretized finite element equations.

There are several algorithms used to construct the Delaunay triangulation [188, 353, 514]. Most of these are based on the Bowyer-Watson algorithm [50, 557], which can be briefly summarized as follows (3D interpretation is given in parentheses):

STEP 1 Define the convex hull within all points will lie. This can be four points (eight points) subdivided into two triangles (five tetrahedra).
STEP 2 Introduce a new point $\boldsymbol{x}_{n+1}$ within the convex hull.
STEP 3 Find all triangles (tetrahedra) whose circumcircle (circumsphere) contains $\boldsymbol{x}_{n+1}$. These identify the triangles (tetrahedra) that will be deleted. A void of elements results.
STEP 4 Find all points belonging to these triangles (tetrahedra).
STEP 5 Find all external edges (faces) of the void resulting from the deletion.

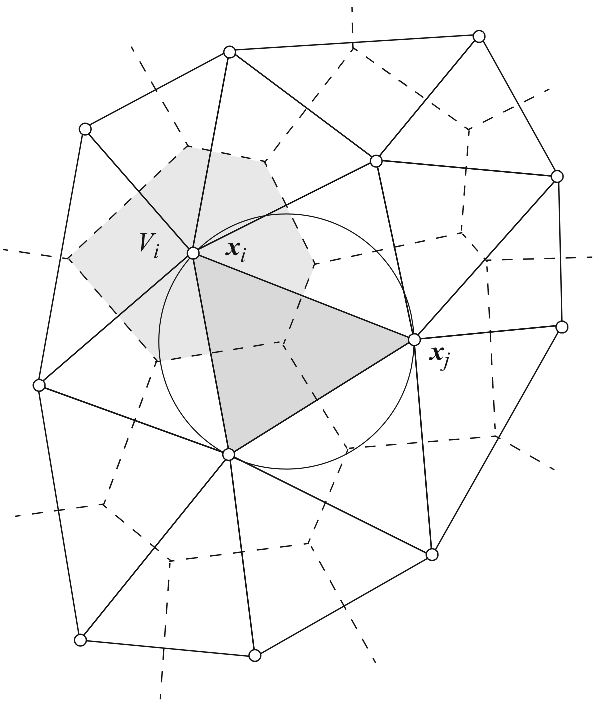


Fig. 15.12 Delaunay triangulation (*solid line*) with circumcircle of a selected triangle and Voronoï regions (*dashed line*)

STEP 6 Form new triangles (tetrahedra) by connecting the found external edges (faces) to the new point $\boldsymbol{x}_{n+1}$.

STEP 7 Add the new elements and the point; update data structure.

STEP 8 Repeat STEPS 2–7 for the next point.

This algorithm provides the basis for unstructured meshing methods. Appropriate treatment is required for recovering points and edges (faces) describing the discretized boundary. An example of Delaunay triangularization is shown in Fig. 15.13 for the circular 2D domain already used for AFT mesh generation in Fig. 15.11 with a similar resolution. We recognize that quite different unstructured meshes can result possessing high irregularity. In comparison to AFT, Delaunay meshes usually look more 'ragged' due to the connectivity of points which is completely free and only constrained via the in-circle criterion.

In the Delaunay mesh generation process the reliable and fast check of the in-circle criterion is a key issue. It states that a point $\boldsymbol{x}_p$ is within the radius R of a circle (sphere) centered at $\boldsymbol{x}_c$ if

$$d_p^2 = (\boldsymbol{x}_p - \boldsymbol{x}_c) \cdot (\boldsymbol{x}_p - \boldsymbol{x}_c) < R^2 \tag{15.7}$$

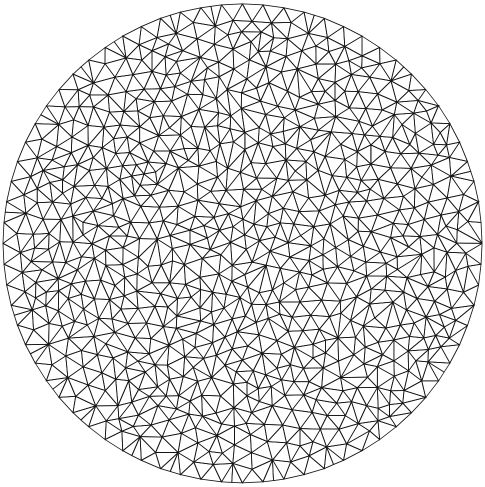


Fig. 15.13 Delaunay triangularization generated for the circular 2D domain of Fig. 15.11 using a homogeneous mesh density

The in-circle check can fail and the triangulation process breaks down if $|d_p - R|^2$ becomes of the order of the computer round-off. To overcome this weakness the test should be better conducted as [353]

$$|d_p - R|^2 < \epsilon_R \tag{15.8}$$

where ϵ_R is a pre-set tolerance that depends on the floating point accuracy of the computer (e.g., 10^{-12} in double precision arithmetic). Another related difficulty in the Delaunay-based meshing may arise for so-called sliver elements. For example, giving the circumcircle for a triangle as

$$(\boldsymbol{x}_i - \boldsymbol{x}_c) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_c) < R^2, \quad i = 1, 2, 3 \tag{15.9}$$

it may happen that the three nodes i lie on a straight line so that $R \to \infty$. In such a case, the point to be inserted has to be rejected and stored for a later use (skip and retry). The in-circle criterion can also break down for certain degenerate point distributions. A common degeneracy arises when the points are distributed in a regular manner. For instance, if in 2D four or more points lie on a circle, the triangulation is no more unique. To overcome this problem, the decision as to whether a point is inside or outside the circumcircle must be consistent for all the triangles involved.

It is possible to define line and point sources to provide an appropriate control of local mesh point spacing during the unstructured meshing process. A useful gradation function has the form [514]

$$\delta(x) = \begin{cases} \delta_1 & \text{if } x \leq x_c \\ \delta_1 e^{|\frac{x-x_c}{D-x_c}| \log 2} & \text{if } x > x_c \end{cases} \tag{15.10}$$

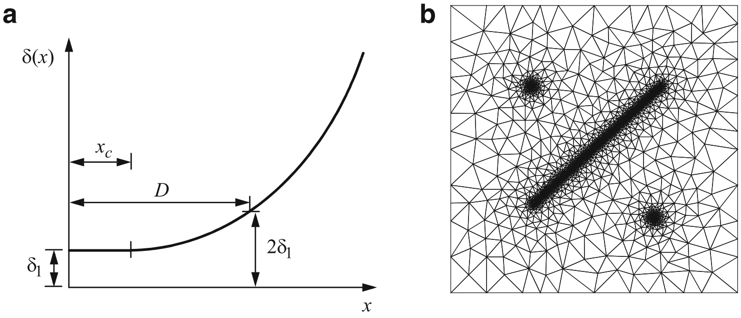


Fig. 15.14 (**a**) Gradation function (15.10) and (**b**) resulting point density of a mesh controlled through point and line sources

to control the target size $\delta(x)$ of an element in the distance x from the point or line source by specifying the quantities δ_1, D and x_c (Fig. 15.14a). An example of a 2D Delaunay mesh which is locally densified around point and line sources through function (15.10) is shown in Fig. 15.14b.

Many DVM-based meshing codes have been developed [514]. For example, a fast and powerful non-commercial 2D triangulator is TRIANGLE [475] which generates high-quality triangular meshes suitable for finite element modeling in many applications.

15.1.3 Mesh Quality Enhancement

Mesh quality represents a rather general term and refers to two aspects: (1) the geometric quality of a mesh which means that the mesh should have a good regularity indicating smoothly varying element sizes having a proper ratio of the maximum to minimum side lengths per element and a reasonable shape (e.g., no excessively deformed and skewed elements, triangles or tetrahedra should not contain large obtuse angles, the minimum angle should be large, number of elements surrounding a nodal point should be limited to avoid rosette-shaped patches), (2) the mesh should be in accordance with the underlying physical problem to be solved, i.e., finer meshes are required where large gradients in the solution variable(s) have to be captured or where discontinuities in material properties or BC's occur; coarser meshes are suited in the far field or in parts of the domain where it can be expected *a priori* that the solution(s) will not vary significantly; thinly shaped (prismatic) elements are useful to adapt properly anisotropic behavior or layered structures. In the following we consider important ways to *a posteriori* improve a generated mesh for a better quality or to better adapt it to a solution.

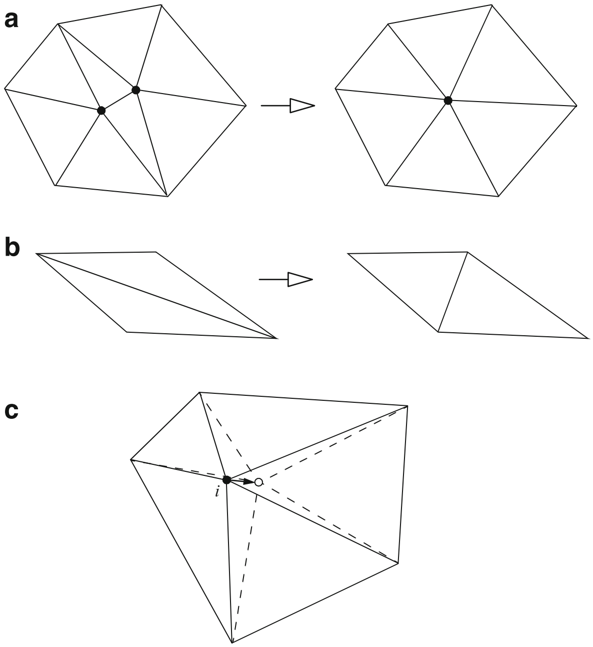


Fig. 15.15 (**a**) Element removal by edge collapse, (**b**) diagonal swapping to maximize the smallest angle in the triangles and (**c**) selective mesh movement of patch node *i*

15.1.3.1 Element Removal, Diagonal Swapping and Selective Mesh Movement

A simple way to eliminate badly deformed elements is in their removal by edge collapsing so that nodes coincide as shown in Fig. 15.15a. The removal leads to a better shaped, more regular element distribution. For meshes of triangular (or tetrahedral) elements local diagonal swapping represents another straightforward procedure performed on a pair of adjacent elements to improve mesh regularity. A 2D example is shown in Fig. 15.15b, where the connectivity of the triangles is changed to eliminate obtuse angles. It attempts to attain improved mesh configurations containing elements with the largest minimum angle. By moving a node i of an element patch (Fig. 15.15c) the mesh quality can be improved in terms of the ratio of edge lengths or the largest minimum angle. The selective movement of i is only allowed while the adjacent elements do not produce singular Jacobians occurring in the coordinate transformations of the elements (cf. Sect. 8.11).

15.1.3.2 Mesh Smoothing

Mesh smoothing technique can be useful to enhance mesh quality. Most typically, it seeks to reposition mesh nodes such that each internal node is at the centroid of the polygon (polyhedron) formed by its adjacent elements. This repositioning is usually done iteratively. In addition, mesh smoothing can be associated with physical items to include solution gradients into the repositioning process. Our preferred approach is the weighted Laplacian smoothing method proposed by Marchant and Weatherill [372]. Let 0 be an internal node which is surrounded by M nodal points of an element patch, its moved coordinates $\boldsymbol{x}_0^{\tau+1}$ at the new relaxation (iteration) level $\tau + 1$ are obtained according to

$$\boldsymbol{x}_0^{\tau+1} = \boldsymbol{x}_0^{\tau} + \omega \frac{\sum_{i=1}^{M} C_{i0}(\boldsymbol{x}_i^{\tau} - \boldsymbol{x}_0^{\tau})}{\sum_{i=1}^{M} C_{i0}} \tag{15.11}$$

where $\boldsymbol{x}_0^{\tau}$ and $\boldsymbol{x}_i^{\tau}$ are the coordinates of node 0 and surrounding nodes $i = 1, \dots, M$, respectively, at the previous iteration τ, ω is the relaxation parameter (usually, set to 0.1) and C_{i0} is a weight factor given by

$$C_{i0} = k_1 + k_2 \left| \frac{\phi_i - \phi_0}{\phi_i + \phi_0} \right| \tag{15.12}$$

in which ϕ represents a solution variable (e.g., hydraulic head, concentration, temperature) and the constants k_1 and k_2 provide a damping to background noise and amplification of gradients, respectively. The iterations (15.11) are performed over all internal nodes of the mesh (excluding nodes belonging to internal boundaries or interfaces of material zones, which should not be moved) and terminate if their coordinate movements fall below a given small tolerance. If the solution dependence is excluded ($\phi_i = \phi_0$), choosing $k_1 = 1$ and taking full relaxation ($\omega = 1$), (15.11) reduces to a simple barycentric smoothing algorithm [188] written in the form

$$\boldsymbol{x}_0^{\tau+1} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{x}_i^{\tau} \tag{15.13}$$

Prototypical examples of mesh smoothing are illustrated in Fig. 15.16.

15.1.4 Prismatic Mesh Topologies

A relatively simple, but highly efficient and robust strategy for generating 3D meshes is the extension of a planar (unstructured or structured) mesh to a third coordinate direction. In such a procedure, each element of a triangular or quadrilateral 2D mesh forms a basis for the construction of layers of pentahedra and hexahedra, respectively, by prolongation in the third direction (Fig. 15.17). Typically in groundwater modeling, the third direction represents the vertical coordinate direction so that layered structures and geologic strata can be appropriately discretized by prisms. All the more because the horizontal extent can be significantly larger in comparison to the layer thicknesses and very thinly shaped elements must result. For numerical reasons prismatic elements have shown to be best suited to

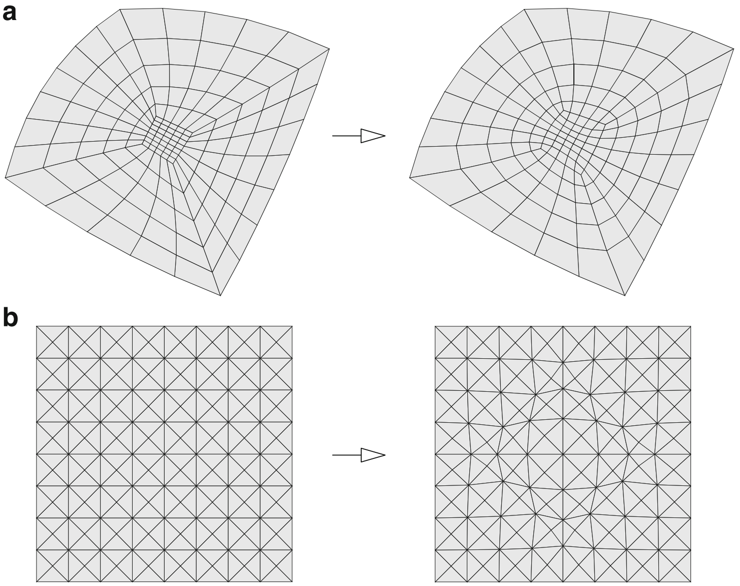


Fig. 15.16 Smoothing of 2D meshes: (**a**) barycentric smoothing of quadrilateral mesh and (**b**) Laplacian smoothing of triangular mesh with solution gradients at central position

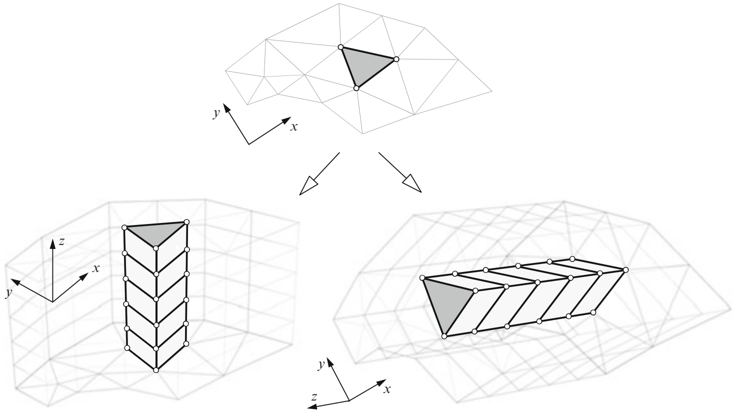


Fig. 15.17 Expanding planar mesh to 3D prismatic mesh by vertical or horizontal prolongation

approximate such type of elongate structures. If the construction of prisms is done exactly in a straight-line direction, we avoid skewed prismatic shapes so that a possible split into tetrahedral elements lead to acceptable angle conditions even for such thin structures. We note that the third direction can also be a horizontal coordinate direction, for instance if we want to expand a vertical 2D mesh into a 3D prismatic topology.

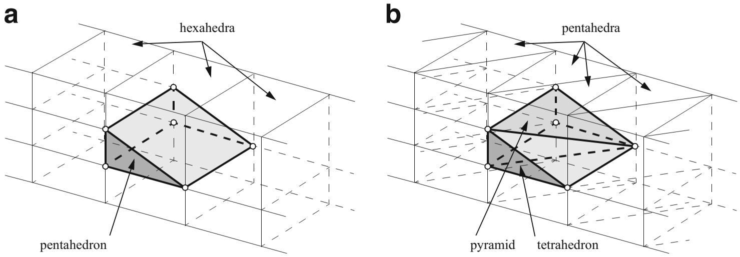


Fig. 15.18 Pinching of (**a**) hexahedral and (**b**) pentahedral topology by inserting transitional elements in form of pentahedra, tetrahedra or pyramids

The resulting prismatic meshes constitute a number of *layers* and *slices*. Layers define the extent of each prism and contain the material properties, while slices define the upper and lower boundaries of each prism and are associated with the nodes of the (linear) prismatic elements to which the computational results are related. Slices can be curved to adapt spatial variation of the layered structures, however, they cannot intersect each other (a typical example is shown in Fig. 9.4). In cases where the number of layers is not present over the whole (horizontal) extent of the domain (e.g., cropped, faulted or eroded layers, local lenses, underground constructions) the prisms can be pinched out to merge layers or pinched in to subdivide layers at a local extent. To retain conformal 3D meshes after pinching, transitional elements in form of pentahedra, tetrahedra and pyramids are used for adapting hexahedral and pentahedral topologies, respectively (Fig. 15.18).

15.1.5 Mesh Refinement and Derefinement

To improve the accuracy of a finite element solution a refined mesh over the computational domain can be useful. However, because mesh refinement increases the computational cost, a selective refinement over the critical areas of the domain would be more efficient. On the other hand, in regions where the solutions have shown sufficiently accurate, the mesh can be possibly derefined to improve the efficiency of the finite element solution. The mesh refinement/derefinement (enrichment/coarsening) can be either performed by a simple trial and error procedure or via a fully automatic mesh adaptation based on error estimation of the numerical solution (see following Sect. 15.1.6).

To refine a given mesh *bisection* of adjacent element edges (faces) is most common. In 2D a quadrilateral element splits into four quadrilaterals and a triangular element splits into four smaller triangles, in 3D a brick element splits into eight hexahedra and a tetrahedral element splits into six smaller tetrahedra as illustrated in Fig. 15.19. For tetrahedral elements various splitting alternatives exist, e.g., [64]. Prismatic hexahedral and pentahedral elements can be suitably split into a number

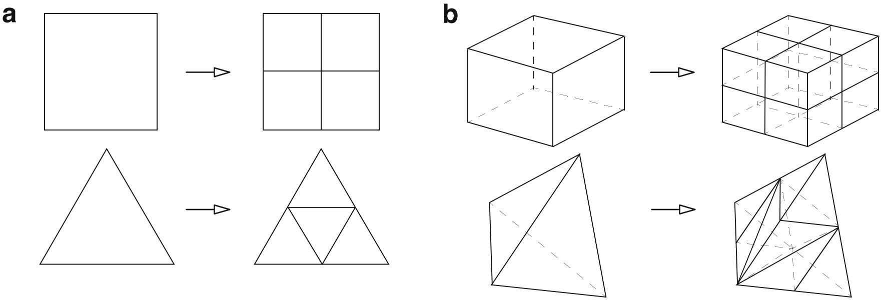


Fig. 15.19 Uniform bisection of (**a**) quadrilateral and triangular element subdivided into four geometrically similar elements and (**b**) hexahedral and tetrahedral element subdivided into eight and six elements of the same type, respectively

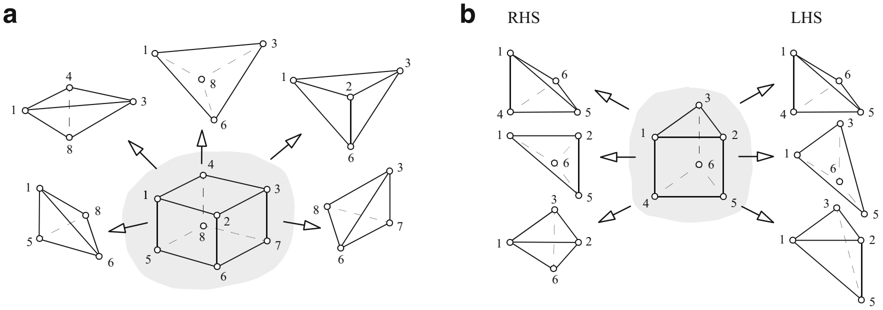


Fig. 15.20 Splitting (**a**) a hexahedron into five tetrahedra and (**b**) a pentahedron into three tetrahedra (each for the RHS and LHS variants)

of tetrahedra (see Fig. 15.20), which can afterwards further be subdivided into tetrahedral elements via bisection. For structured quadrilateral (brick-type) meshes bisection requires mesh lines running to the outer boundaries to retain the four-neighbor (eight-neighbor) structure, similar to grids used in FDM, suitable at most for a global refinement in FEM. To refine locally quadrilateral (brick-type) meshes transitional quadrilaterals become necessary to create conformal all-quadrilateral meshes (Fig. 15.21), however, their use is limited and can lead to badly shaped elements when the local refinement is consecutively applied.

Much more flexibility is provided by using unstructured triangular (pentahedral or tetrahedral) element meshes, which are usually preferred for purposes of adaptive and local mesh refinement/derefinement. We start from a given triangulation (the basic mesh) obtained by using a mesh generator such as described above (Sect. 15.1.2). A powerful and highly flexible refinement/derefinement strategy has been proposed by Bank [23, 24] which is called *red-green triangulation*. It uses regular bisection of a triangular element into four geometrically similar triangles. Such bisected elements are called *red*. To retain conformal meshes (i.e., no hanging nodes) adjacent elements sharing a bisected edge of a red element are irregularly

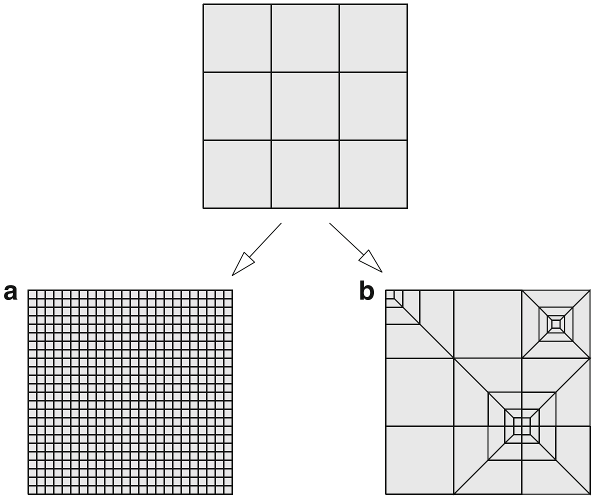


Fig. 15.21 Regular 3 × 3 quadrilateral mesh being three times consecutively refined by (**a**) global refinement leading to a uniform $3 \cdot 2^3 \times 3 \cdot 2^3$ mesh resolution and (**b**) local refinement around selected patch nodes and at a single element

subdivided into two triangles. They are called *green* elements (Fig. 15.22). Green elements, however, have a useful property: once a refinement of green elements is required afterwards, their irregular two-triangle subdivision is abolished and replaced by a regular one-to-four triangle split, so that the green element turns to a red element (Fig. 15.23). This procedure keeps the angles of the refined triangles in acceptable bounds and mesh quality is retained even in hierarchies of element refinements. The process can be repeatedly performed over an arbitrary number of refinement *levels* so that red elements can be further subdivided in a systematic hierarchical manner (Fig. 15.24). In doing so, elements introduced by refinement are regarded as offspring of coarser parent elements and a derefinement (coarsening) of an already refined mesh becomes easy by turning back within the refinement hierarchy, where elements nested within a parent element are removed and the parent is restored as the element (Fig. 15.25). Note, however, coarsening beyond the basic mesh (refinement level 0) is not possible. Bank's red-green triangulation can easily be applied to prismatic pentahedral 3D meshes when accepting the refinement/derefinement procedure for each layer. Alternative refinement techniques have been developed for tetrahedral meshes, e.g., [347, 353, 501].

15.1.6 Adaptive Mesh Refinement (AMR)

A successful use of numerical methods (FEM, FVM) requires significant expertise and cost. The accuracy and reliability of the computations are of theoretical as well as of practical interest and represent a central problem in the numerical anal-

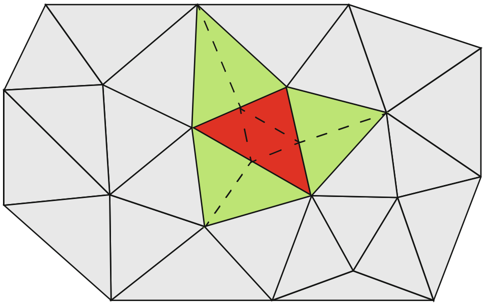


Fig. 15.22 Bank's red-green triangulation: regular bisection of *red element* into four triangles and adjacent *green elements* each irregularly subdivided into two triangles (*shown dashed*)

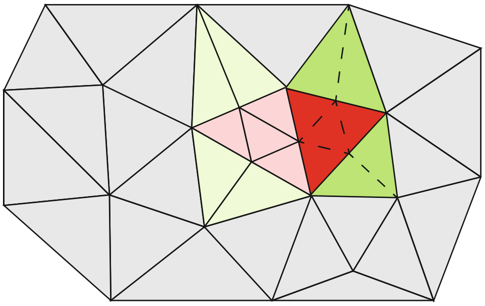


Fig. 15.23 Regular refinement of a *green triangle* of the mesh shown in Fig. 15.22 to avoid a second bisection of irregular triangles. New refinements are shown as *dashed lines*

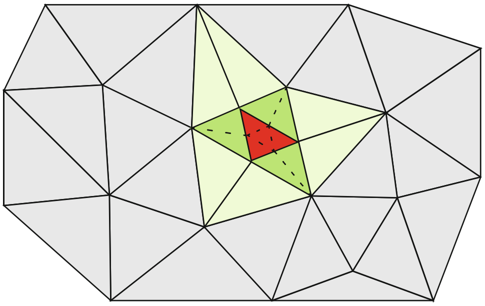


Fig. 15.24 Second refinement (level 2) of the *red triangle* of the refined mesh (at level 1) shown in Fig. 15.22. New refinements are shown as *dashed lines*

ysis [16]. Promising self-adaptive strategies for an automatic quantitative control of the discretization error(s) have been developed [353, 433, 545, 590]. Ideally, an adaptive method should be *reliable* in the sense that the discretization-error control should be guaranteed, and also efficient in the sense that the computational effort should remain within acceptable bounds. However, for the most practically relevant problems analytical (a priori) error estimates do not exist. Nevertheless, adaptive finite element algorithms can be constructed on the basis of *a posteriori* error estimates. They postulate that a finite element solution can indicate which regions in a given domain need refinement or allow coarsening.

An adaptive procedure is divided into two phases: error estimation and mesh refinement/derefinement. Since the exact solution is not known, a method to approximate the error of a numerical solution is needed. Thus, an appropriate

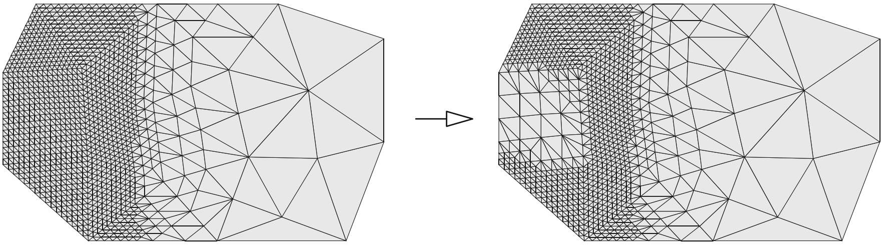


Fig. 15.25 Multiple refinement applied to the mesh shown in Fig. 15.22 and attained derefinement in a selected area

measure, a norm, or seminorm, of the error has to be defined (cf. Sect. 8.4.2). Commonly, the error energy norm

$$\|\boldsymbol{e}\|^2 = \int_\Omega \boldsymbol{e}^T \mathcal{L}(\boldsymbol{e}) d\Omega \tag{15.14}$$

is used [593], where $\mathcal{L}$ is a differential operator and $\boldsymbol{e}$ is a local error measure. Denoting the exact and the approximate finite element solution by $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$, respectively, $\boldsymbol{e}$ is defined as

$$\boldsymbol{e} = \boldsymbol{\phi} - \hat{\boldsymbol{\phi}} \tag{15.15}$$

For advection-dispersion equations we can determine an equivalent energy functional if the differential operator $\mathcal{L}$ in (15.14) is expressed by the dispersive flux vector $\boldsymbol{f}^d$ from (8.2), viz.,

$$\mathcal{L}(\boldsymbol{e}) = -\nabla \cdot (\boldsymbol{f}^d - \hat{\boldsymbol{f}}^d) \tag{15.16}$$

Inserting (15.15) and (15.16) into (15.14), integrating by parts, using Gauss's integral theorem (2.77) and noting that the error vanishes on the boundary Γ because $\boldsymbol{\phi}$ or $\boldsymbol{f}^d$ are prescribed there, the error estimator becomes[3]

[3] Alternatively, to find an appropriate energy expression for coupled variable-density flow, mass and heat transport in porous media, the internal Clausius-Duhem entropy production $\rho\Upsilon \geq 0$, (3.125), can be utilized. It provides a physically consistent functional in which all relevant state variables of the nonlinearly coupled process in form of Darcy flux $\boldsymbol{q}$, hydraulic head h, species mass C_k and temperature T are present. A simplified version of (3.125) yields

$$\bar{\Upsilon}(\boldsymbol{q}, T, C_k) = T\,\rho\Upsilon = \rho_0 g \boldsymbol{q}\cdot(\boldsymbol{K}^{-1}\cdot\boldsymbol{q}) + \frac{1}{T}\big(\nabla T\cdot(\boldsymbol{\Lambda}\cdot\nabla T)\big) + \sum_k \frac{\partial\mu_k}{\partial C_k}\big(\nabla C_k\cdot(\boldsymbol{D}_k\cdot\nabla C_k)\big) \geq 0$$

and the following entropy error norm appears suitable

$$\|\boldsymbol{e}\|^2 = \int_\Omega \bar{\Upsilon}(\boldsymbol{q} - \hat{\boldsymbol{q}}, T - \hat{T}, C_k - \hat{C}_k) d\Omega$$

$$\|\boldsymbol{e}\|^2 = \int_\Omega \big(\nabla(\phi - \hat{\phi})\big) \cdot \big(\boldsymbol{f}^d - \hat{\boldsymbol{f}}^d\big) d\Omega \tag{15.17}$$

where the dispersive fluxes $\boldsymbol{f}^d$ and $\hat{\boldsymbol{f}}^d$ are expressed by their exact and approximate finite element solutions according to (8.2), which depend on the state variable ϕ and $\hat{\phi}$, respectively. In fact, (15.17) forms a functional of the hydraulic head h, species mass C_k and temperature T variables.

While (5.17) characterizes the error, a measure for the exact solution is also needed. Such a norm can be similarly derived as

$$\|\phi\|^2 = \int_\Omega \phi^T \mathcal{L}(\phi) d\Omega = \int_\Omega (\nabla\phi) \cdot \boldsymbol{f}^d d\Omega \tag{15.18}$$

We recognize from (15.17) and (15.18) that exact values of hydraulic head, species mass and temperature gradients (derivatives), which are generally unknown, must be required. Therefore, these values must be approximated. A usual way is to compute these derivatives by a higher-order approximation. Fluxes of higher-order accuracy can be easily obtained by using global or local smoothing techniques, such as already discussed in Sects. 8.19.1.1 and 8.19.1.2, respectively. They can also be obtained by adequate SPR technique, as described in Sect. 8.19.1.3. In doing so, we compute the higher accurate and continuous $\nabla\phi$ by a flux smoothing or recovery technique, while the lower accurate and discontinuous $\nabla\hat{\phi}$ is obtained by direct differentiation similar to (8.402). The norms $\|\boldsymbol{e}\|$ and $\|\phi\|$ are evaluated as the sum of their respective element contributions, viz.,

$$\|\boldsymbol{e}\|^2 = \sum_{e=1}^{N_E} (\|\boldsymbol{e}\|^e)^2, \quad \|\phi\|^2 = \sum_{e=1}^{N_E} (\|\phi\|^e)^2 \tag{15.19}$$

where the element norms $\|\boldsymbol{e}\|^e$ and $\|\phi\|^e$ are given for each element e as

$$\begin{aligned} (\|\boldsymbol{e}\|^e)^2 &= \textstyle\int_{\Omega^e} \big(\nabla(\phi^e - \hat{\phi}^e)\big) \cdot \big(\boldsymbol{f}^{d^e} - \hat{\boldsymbol{f}}^{d^e}\big) d\Omega^e \\ (\|\phi\|^e)^2 &= \textstyle\int_{\Omega^e} (\nabla\phi^e) \cdot \boldsymbol{f}^{d^e} d\Omega^e \end{aligned} \tag{15.20}$$

equivalent to (15.17) and (15.18), respectively. Note that only the square of the error norm is additive.

Usually, global and local error criteria can be defined to rate the accuracy of the solution by using appropriate relative quantities, where a mesh is considered *optimal* if these global and/or local (elementwise) error criteria are satisfied [402]:

(a) *Global error condition* requires that the error relative to the exact solution must be smaller than a permissible error tolerance, such as

where $\boldsymbol{q}$, C_k, T are the exact solutions and $\hat{\boldsymbol{q}}$, $\hat{C}_k$, $\hat{T}$ are the approximate finite element solutions. The Darcy flux $\boldsymbol{q} = \boldsymbol{q}(h, C_k, T)$ or $\hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}(\hat{h}, \hat{C}_k, \hat{T})$ takes the form of (11.1).

$$\varsigma = \frac{\|\boldsymbol{e}\|}{\|\boldsymbol{\phi}\|} \le \epsilon_\Delta \tag{15.21}$$

where ϵ_Δ represents the spatial AMR-specific error tolerance to be set. A *global error indicator* ς^g is defined as

$$\varsigma^g = \frac{\|\boldsymbol{e}\|}{\epsilon_\Delta \|\boldsymbol{\phi}\|} \tag{15.22}$$

so that $\varsigma^g \le 1$ satisfies the global error criterion, whereas $\varsigma^g > 1$ indicates a need for further mesh refinement.

(b) *Local error condition* is used to bound the error in each element of a mesh. It can be expressed as

$$\|\boldsymbol{e}\|^e = \|\boldsymbol{e}\|_r^e, \quad (e = 1, \ldots, N_\text{E}) \tag{15.23}$$

where $\|\boldsymbol{e}\|_r^e$ is a 'required' error norm valid for the element. A *local error indicator* ς^e is defined as

$$\varsigma^e = \frac{\|\boldsymbol{e}\|^e}{\|\boldsymbol{e}\|_r^e} \tag{15.24}$$

so that $\varsigma^e = 1$ indicates an *optimal* size of element e, whereas $\varsigma^e > 1$ and $\varsigma^e < 1$ indicate that the size of element e needs further refinement or derefinement, respectively. Clearly, the definition of $\|\boldsymbol{e}\|_r^e$ is a key issue, which strongly affects the mesh resolution and element size distribution. The following two major AMR strategies will specify different mesh optimality criteria for $\|\boldsymbol{e}\|_r^e$.

In the preferred AMR process both global *and* local error conditions will be satisfied. To control the mesh refinement and derefinement the following *element refinement parameter* is used

$$\xi^e = \varsigma^e \varsigma^g = \frac{\|\boldsymbol{e}\| \|\boldsymbol{e}\|^e}{\epsilon_\Delta \|\boldsymbol{\phi}\| \|\boldsymbol{e}\|_r^e} \tag{15.25}$$

which results from combining (15.22) and (15.24). It represents an appropriate indicator to relate the actual element error to the distributed value of the permissible error over the mesh. Thus, $\xi^e \ge 1$ will indicate that element e needs further refinement, whereas $\xi^e < 1$ indicates that both the local and global error conditions are satisfied in element e. We employ the following two different optimality criteria in the AMR process:

(A) Zienkiewicz and Zhu's optimality criterion [594]. By using this criterion the global error is equally distributed for all elements N_E of a mesh. The 'required' error for each element is considered as an average of the global error per element e, i.e.,

$$(\|\boldsymbol{e}\|_r^e)^2 = \frac{\|\boldsymbol{e}\|^2}{N_{\mathrm{E}}} \tag{15.26}$$

Then, the element refinement parameter ξ^e (15.25) gives

$$\xi^e = \frac{\|\boldsymbol{e}\|^e \sqrt{N_{\mathrm{E}}}}{\epsilon_\Delta \|\phi\|} \tag{15.27}$$

(B) Oñate and Bugeda's optimality criterion [402]. This criterion distributes the specific error in form of the square of the error per unit area (or volume) over the whole mesh, viz.,

$$\frac{(\|\boldsymbol{e}\|_r^e)^2}{\Omega^e} = \frac{\|\boldsymbol{e}\|^2}{\Omega} \tag{15.28}$$

so that the element refinement parameter ξ^e (15.25) gives

$$\xi^e = \frac{\|\boldsymbol{e}\|^e}{\epsilon_\Delta \|\phi\|} \left(\frac{\Omega}{\Omega^e}\right)^{\frac{1}{2}} \tag{15.29}$$

Note that Zienkiewicz and Zhu's optimality criterion (15.27) and Oñate and Bugeda's optimality criterion (15.29) coincide for meshes consisting of elements equal in size $\frac{\Omega}{\Omega^e} = N_{\mathrm{E}}$. However, for unstructured meshes both criteria are generally different and the advantage of Oñate and Bugeda's optimality criterion is that more and smaller elements are generated in those areas where steep solution gradients have to be captured, so as physically useful.

The element refinement parameter ξ^e, (15.27) or (15.29), indicates elements which are above or below a permissible error measure. To meet the required accuracy according to the chosen optimality criterion, there are three strategies: (i) the $h-$adaptation, where the mesh is refined or coarsened locally, (ii) the $p-$adaptation, in which the polynomial degree of the finite element basis (trial) functions is locally increased or decreased, while the mesh is not changed and (iii) the $r-$adaptation, in which the mesh is relocated or moved. These strategies may be used singly or in combination and can also be combined with mesh smoothing (see Sect. 15.1.3.2).

For the present problem class the $h-$refinement has shown best suited and robust. Basically, the $h-$refinement strategy can be treated hierarchically. The finer mesh is created within the coarser mesh, and vice versa. Triangles in two dimensions and tetrahedra in three dimensions are particularly convenient and flexible in contrast to quadrilateral and hexahedral elements. Most common is the red-green triangulation (cf. Sect. 15.1.5 above), in which the basic (coarse) mesh is locally refined and redefined in accordance with the error criteria to be satisfied for each element of the

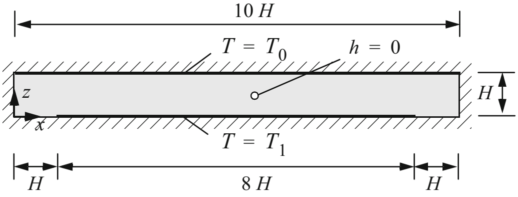


Fig. 15.26 Definition of Elder's long-heater problem – domain and BC's

mesh. Practically, whenever the element refinement parameter $\xi^e \geq 1$, the element e is indicated as *red* and becomes refined by a regular one-to-four triangle split; otherwise if $\xi^e < 1$ the element e can be potentially coarsened. To avoid extremely small element sizes in critical regions, the refinement is terminated if the element area (volume) falls below a minimum magnitude. This refinement/derefinement process is iteratively repeated until the mesh becomes *optimal* in a sense that ξ^e will be approximately unity for all elements of the mesh and no refinement is necessary within the AMR-specific error tolerance pre-set by ϵ_Δ.

For transient problems the spatial adaptation has to be embedded in an appropriate time-stepping cycle. In using an automatic time-stepping control, the overall solution becomes fully adaptive for the proposed predictor-corrector technique. However, the computational effort can increase substantially and an efficient monitoring of the temporal and spatial refinement/derefinement is necessary. Ideally, the spatial adaptation is performed at each time plane $n + 1$. Alternatively, one can check the error criteria only after a certain step number of advances in time and allow remeshing only if error variation with time is significant [169].

The AMR procedure in combination with adaptive time stepping monitors the actual spatial and temporal discretization requirements ('watching physics'). We illustrate the use of such a fully adaptive solution procedure for the solution of Elder's variable-density 'long-heater problem' [154] as sketched in Fig. 15.26. The used simulation parameters, IC and BC's are summarized in Table 15.1. The assumed flow and thermal parameters are isotropic and constant, thermodispersivity does not occur. At the center of the cross-sectional domain, measuring $10H \times H$ units, a reference hydraulic head h is set to zero. On the top boundary a constant temperature T_0 is imposed, while at the bottom of the domain a higher temperature $T_1 > T_0$ is imposed over a distance of $8H$ units. On all the remaining parts of the boundary, natural BC's are set. Boundary values and parameters are equivalent to a thermal Rayleigh number Ra_t, (11.26), of 200.

Applying the AMR strategies to the transient long-heater problem we observe a different effect on the mesh resolution in time depending on the used optimality criterion. Figure 15.27 shows the sequence of refined meshes with related results of isotherms and streamline patterns at selected (dimensionless) times $\hat{t}$ by using Oñate and Bugeda's optimality criterion. The equivalent results for the Zienkiewicz and Zhu's optimality criterion are shown in Fig. 15.28. In both cases, the simulation

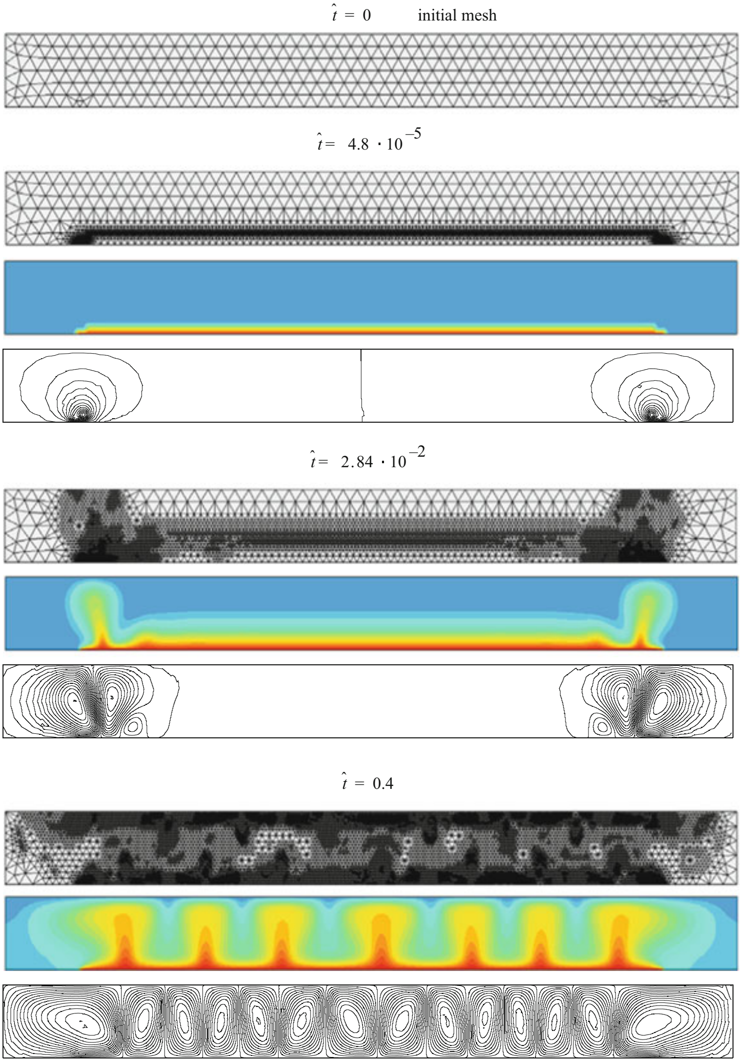
Table 15.1 Parameters and conditions used for Elder's long-heater problem

Quantity	Symbol	Value	Unit
Study domain shown in Fig. 15.26.			
Domain measure (height; length)	H; $10H$		
Thermal Rayleigh number, (11.26)	Ra_t	200	1
Specific storage coefficient	S_o	0	m^{-1}
Thermal storage coefficient, (11.21)	S_λ	1	1
Flow BC			
Dirichlet-type BC at center	$h(x,z) = h_D(5H, \frac{H}{2})$	0	m
Dimensionless temperature[a] *IC and BC's*			
Initial condition (IC) of temperature	$\hat{T}_0$	0	1
Dirichlet-type BC at top ($z = H$)	$\hat{T}_D \quad (0 \leq x \leq 10H)$	0	1
Dirichlet-type BC at bottom ($z = 0$)	$\hat{T}_D \quad (1H \leq x \leq 9H)$	1	1
FEM			
Adaptive triangular meshes, GFEM (no upwind), OB approximation			
Initial time step size[b]	$\Delta\hat{t}_0$	$1.33 \cdot 10^{-6}$	1
Temporal RMS error tolerance (AB/TR)	ϵ	10^{-3}	1
Spatial AMR error tolerance (Zienkiewicz-Zhu)	ϵ_Δ	$5 \cdot 10^{-2}$	1
Spatial AMR error tolerance (Oñate-Bugeda)	ϵ_Δ	$3 \cdot 10^{-3}$	1
Simulation time period[b]	$\hat{t}_{\text{end}}$	0.4	1

[a] Dimensionless temperature $\hat{T} = \frac{T}{T_1 - T_0}$, where T_0 and $T_1 > T_0$ occur at the top and bottom boundary, respectively

[b] Dimensionless time $\hat{t}$ as defined in (11.23)

starts from a coarse mesh consisting of only 658 triangles. We can see how the thermal gradients are appropriately adapted by the mesh varying in time. The distribution of elements, the taken refinement levels and the trailing derefinement of the mesh have shown very dynamic during the evolution of the thermal convection process. Zienkiewicz and Zhu's strategy leads to meshes exhibiting more spread of refinement around gradients, while Oñate and Bugeda's strategy enriches the mesh more locally in dependence on the evolving temperature gradients. Both AMR strategies take about 900 adaptive time steps of variable size by using the AB/TR predictor-corrector time integrator showing similar characteristics in the time step behavior (Fig. 15.29a), however, their resulting mesh resolutions develop rather differently as depicted in Fig. 15.29b for the total number of elements N_{E} in time. At later times more refined meshes are obtained from Zienkiewicz and Zhu's AMR strategy. However, we can recognize from Figs. 15.27 and 15.28 the different meshes have no or only very small influence on the achieved solution results. Furthermore, we note that the AMR procedures do not lead to actually symmetric meshes, which may imply additional perturbations in the convection process. In Figs. 15.29c and 15.29d the statistical evaluation gives additional insight into the spatial and temporal behavior of the element refinement parameter ξ^e in form of its average and mean deviation. Since the AMR process has been performed at each time plane, the mesh tends to slightly fluctuate in time between refinement and derefinement once ξ^e approaches unity (i.e., reaching the optimal state). The suitable choice of



$\hat{t} = 0$ initial mesh

$\hat{t} = 4.8 \cdot 10^{-5}$

$\hat{t} = 2.84 \cdot 10^{-2}$

$\hat{t} = 0.4$

Fig. 15.27 AMR simulation results of Elder's long-heater problem for thermal Rayleigh number $\mathrm{Ra}_t = 200$ at different dimensionless times $\hat{t}$ based on Oñate and Bugeda's optimality criterion (15.29): adapted meshes, fringed isotherms and contoured streamlines

an additional threshold parameter could remedy these transient mesh oscillations. In combination with the adaptive time stepping, the overall AMR procedure is prone to underestimate ξ^e, which leads to potentially '*superoptimal*' meshes with $\xi^e < 1$ indicating more refinement than necessary. Practically, as revealed in Fig. 15.29c both automatic AMR strategies realizes an average value of element refinement parameter of $\overline{\xi^a} \approx 0.7$ over the complete time range (where Oñate and Bugeda's

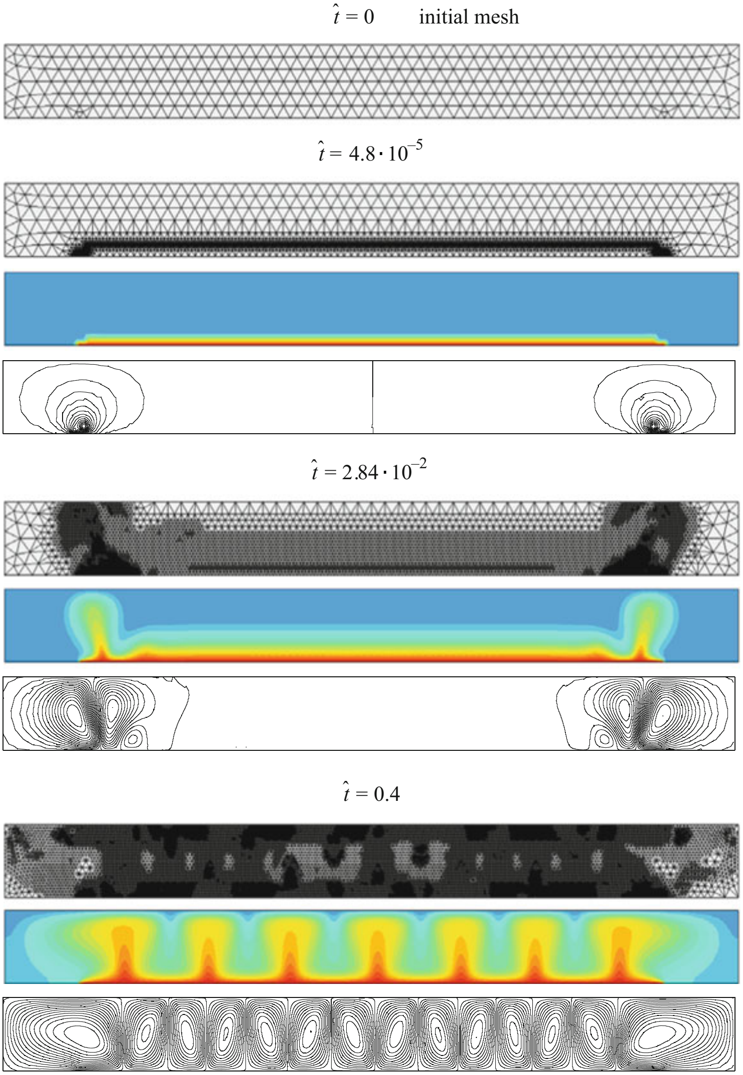

$\hat{t} = 0$ initial mesh
$\hat{t} = 4.8 \cdot 10^{-5}$
$\hat{t} = 2.84 \cdot 10^{-2}$
$\hat{t} = 0.4$

Fig. 15.28 AMR simulation results of Elder's long-heater problem for thermal Rayleigh number $\mathrm{Ra}_t = 200$ at different dimensionless times $\hat{t}$ based on Zienkiewicz and Zhu's optimality criterion (15.27): adapted meshes, fringed isotherms and contoured streamlines

strategy looks somewhat better), except at beginning when the AMR process is initiated in the adaptive time stepping, which starts from a very coarse mesh. As a consequence of the transient AMR control, the mean deviation of element refinement parameter $\overline{\xi^{\sigma}}$ as shown in Fig. 15.29d could not run lower than about 0.3 over the most time stages. In conclusion, taking into account that the overall AMR solution process is controlled by only two error tolerance measures (ϵ for

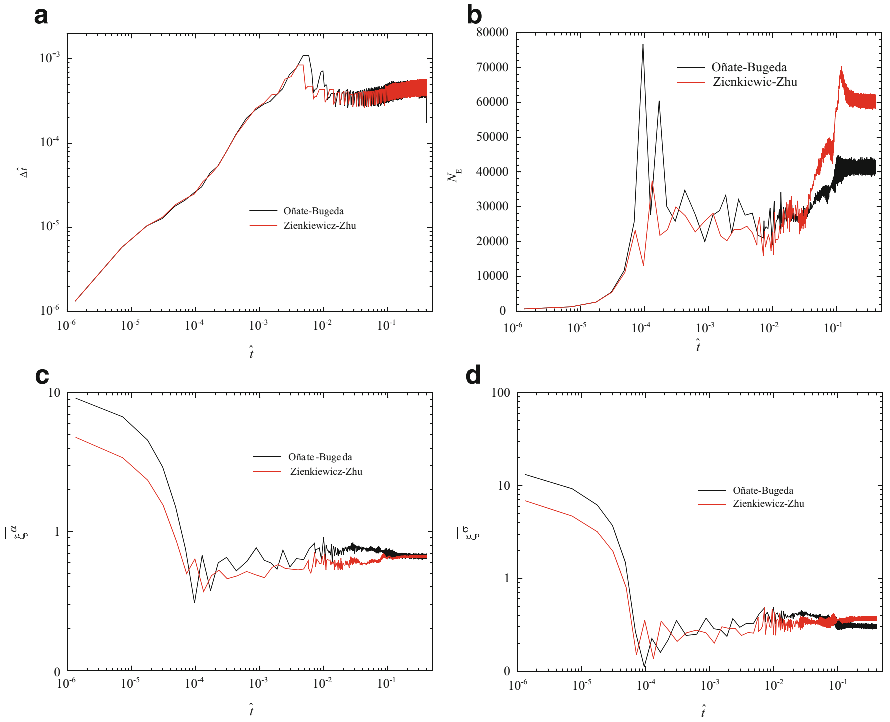


Fig. 15.29 Characteristic results for the AMR solution of Elder's long-heater problem in comparison between Oñate and Bugeda's and Zienkiewicz and Zhu's optimality criterion: History of (**a**) the adaptive time steps $\Delta\hat{t}$, (**b**) the number of total elements N_{E} (mesh resolution), (**c**) the average value of element refinement parameter $\overline{\xi^a} = \frac{1}{N_{\mathrm{E}}}\sum_e \xi^e$ and (**d**) the mean deviation of element refinement parameter $\overline{\xi^\sigma} = \frac{1}{N_{\mathrm{E}}}\sum_e |\xi^e - \overline{\xi^a}|$

the adaptive time stepping and ϵ_Δ for the spatial AMR control), the proposed fully adaptive strategy has shown promising in reducing 'user-defined' discretization influences from the finite element solutions.

15.2 Particle Tracking Techniques

15.2.1 General

Pathlines are widely used to describe and visualize fluid motion (cf. definitions introduced in Sect. 2.1.11). They indicate the advective movement of massless particles in a known velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$ that are introduced at suitable points in the flow. The trace of a swarm of particles provides visual information on flow direction.

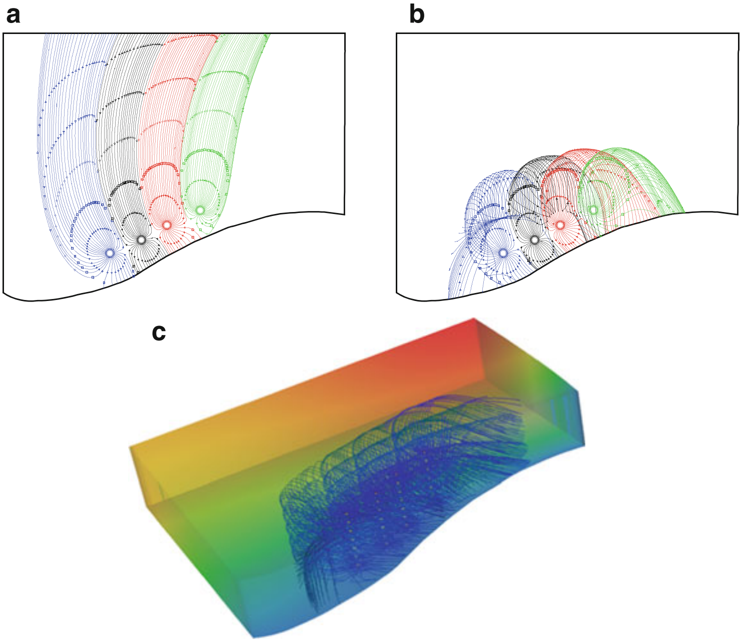


Fig. 15.30 Pathlines and isochrons for (**a**) 2D steady-state flow field, (**b**) 2D unsteady flow field and (**c**) 3D unsteady flow field computed via backward particle tracking around pumping wells

Pathlines can be marked at regular time intervals to feature travel times in form of isochrons. A typical example is shown in Fig. 15.30 for steady-state and transient flow caused by a number of pumping wells in an unconfined aquifer. We note that pathlines in steady-state flow coincide with streamlines (Sec. 2.1.11). While streamlines do not intersect (Fig. 15.30a, except at a stagnation point where $\boldsymbol{v} \equiv \mathbf{0}$), pathlines can intersect themselves or other pathlines (Fig. 15.30b).

Particle paths are governed by the set of ODE, (2.96)

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{v}(\boldsymbol{x}, t), \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{15.30}$$

where $\boldsymbol{x}_0$ is the initial location of the particle. The location $\boldsymbol{x}$ and the velocity $\boldsymbol{v}$ can be given in Cartesian or cylindrical coordinates, (2.70) and (2.71), respectively. The initial value problem (15.30) must be solved to describe the particle trajectory for the known velocity field $\boldsymbol{v}$, viz.,

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{v}\big(\boldsymbol{x}(t), t\big)\, dt \tag{15.31}$$

where $\boldsymbol{x}(t)$ corresponds to the position of the particle at time t. The integration (15.31) is known as *particle tracking*, for which a number of different approaches exist, e.g., [222, 424, 587]. It is performed in a postprocessing computation for which the velocity field is available at a series of discrete points and time planes. For a sufficiently simple velocity $\boldsymbol{v}$, (15.31) may even be integrated exactly in an analytical (semianalytical) tracking procedure. More general and favorable, however, are numerical tracking methods, commonly preferred in FEM. In the following, we shall describe the basic principles of both tracking strategies to show their advantages and drawbacks in the present modeling context.

Independently of the actually used particle tracking solution strategy, both (semi)-analytical and numerical methods, there are a number of sources of errors, which influence the accuracy of any particle tracking solution [424]:

1. *Errors in nodal velocities.* The accuracy of the derived velocity distribution computed from the flow model is most crucial. It depends on the spatial and temporal resolution and the used techniques to derive velocities (as a secondary variable) from primary solution variables.
2. *Spatial interpolation.* Even if the nodal (discrete) velocities are exact, their interpolation, needed for computing the velocities at any position in the computational domain, affects the results of tracking.
3. *Error of numerical integration.* The solution of the governing ODE (15.30) can introduce errors. While analytical methods lead (potentially) to an exact solution, numerical methods can cause errors due to truncation errors and round-off errors. However, by using higher-order numerical approaches with fully adaptive error control, these sources of errors can be kept sufficiently small.
4. *Exit error.* This error can be introduced in determining the position of particles when they leave discrete units (elements, cells) of the computational domain.

Obviously, in the integration (15.31) the evaluation of the velocity $\boldsymbol{v}$ is a key issue. From the numerical flow modeling via FEM (or others), the velocity $\boldsymbol{v}$ is only available as discrete values, i.e., at nodal points and time planes. It is obvious from (15.31) that $\boldsymbol{v}$ can only be successfully integrated if the velocity is unique and sufficiently smooth, at least piecewise-smooth. Practically, a *continuous* (smoothed) velocity field $\boldsymbol{v}$ is required which can be obtained by using the smoothing techniques as described in Sects. 8.19.1, 9.7, 10.11, 11.7 and 14.4. The resulting velocity $\boldsymbol{v}(\boldsymbol{x}, t)$ smoothly behaves over the discretized domain $\bar{\Omega}$ so that for any location $\boldsymbol{x} \in \bar{\Omega}$ a unique and piecewise continuous velocity value is available by using interpolation (Sect. 15.4).

To track particles due to pure advection, the velocity $\boldsymbol{v}$ has to be an intrinsic quantity. In porous-media flow, the intrinsic velocity results from evaluating the Darcy equation, e.g., (3.303), (4.38), (9.1), (10.1), (11.1), as the pore velocity (3.240) given by

$$\boldsymbol{v} = \frac{\boldsymbol{q}}{s\,\varepsilon} \tag{15.32}$$

where $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{x}, t)$ is the Darcy velocity, $s = s(\boldsymbol{x}, t) > 0$ is the saturation and $\varepsilon = \varepsilon(\boldsymbol{x}, t) > 0$ is the porosity. To track particles which are subjected in addition to sorption, the advection is represented through the use of a retarded velocity and the intrinsic velocity results as

$$\boldsymbol{v} = \frac{\boldsymbol{q}}{s\, \varepsilon\, \Re_k} \tag{15.33}$$

where $\Re_k = \Re_k(\boldsymbol{x}, t) \geq 1$ is the retardation factor (see Table 3.8) valid for the particle associated with species k. We note that the actual values of the intrinsic velocity are important in presence of a transient velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$ and in determining travel times (isochrons), while for stationary velocity $\boldsymbol{v}(\boldsymbol{x})$ the shape of particle trajectory (pathline, streamline) will not be affected by the 'scaling' operation according to (15.32) or (15.33), i.e., becoming independent of s, ε and $\Re_k$.

The particle tracking procedures can be used either as *forward tracking* or as *backward tracking*. In a forward tracking the velocity $\boldsymbol{v}$ is used in its original values and the particles are commonly started in sources, recharge wells or upgradient areas. On the other hand, in a backward tracking the velocity is applied in a reversed manner so that the particles track in the reverse flow direction starting from sinks, pumping wells or downgradient areas (see Fig. 15.30), viz.,

$$\boldsymbol{v} := \begin{cases} +\boldsymbol{v} & \text{forward tracking} \\ -\boldsymbol{v} & \text{backward tracking} \end{cases} \tag{15.34}$$

15.2.2 Pollock's Semianalytical Tracking Method

In classic finite-difference groundwater modeling Pollock [425] introduced an efficient algorithm for computing pathlines via a semianalytical approach. It is termed a 'semianalytical' method because the velocities are evaluated numerically (actually, on a regular FDM grid) while the pathline integration is performed analytically. Let us describe at first the basic version of Pollock's method. Afterwards, possible extensions and generalizations will be discussed. We consider a block-shaped difference cell with spacing $\Delta x \times \Delta y \times \Delta z$, where on its surfaces six constant velocity values are given (Fig. 15.31). In Pollock's method it must be assumed that the velocity in the cell (1) behaves steady-state and (2) varies linearly. Under these assumptions the velocity is approximated as

$$\boldsymbol{v}(\boldsymbol{x}) = \begin{pmatrix} v_x(x) \\ v_y(y) \\ v_z(z) \end{pmatrix} = \begin{pmatrix} A_x(x - x_1) + v_{x1} \\ A_y(y - y_1) + v_{y1} \\ A_z(z - z_1) + v_{z1} \end{pmatrix}, \quad \begin{aligned} A_x &= (v_{x2} - v_{x1})/\Delta x \\ A_y &= (v_{y2} - v_{y1})/\Delta y \\ A_z &= (v_{z2} - v_{z1})/\Delta z \end{aligned} \tag{15.35}$$

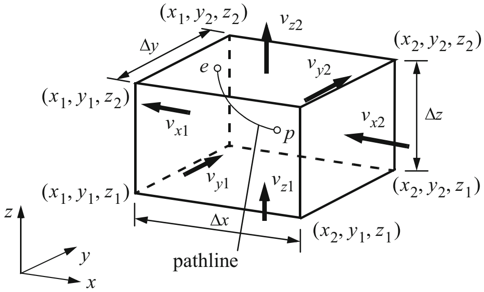


Fig. 15.31 Definition of a 3D difference cell ($\Delta x = x_2 - x_1$, $\Delta y = y_2 - y_1$, $\Delta z = z_2 - z_1$) with its velocity components ($v_{x1}, v_{x2}, v_{y1}, v_{y2}, v_{z1}, v_{z2}$) on the six cell surfaces and a possible particle path between surface entry point p and surface exit point e

and the governing ODE (15.30) of particle tracking takes separated expressions for the $x-$, $y-$ and $z-$components as

$$\begin{array}{lcl} \frac{dx}{dt} = v_x(x), & \text{or} & \frac{dx}{v_x(x)} = dt \\ \frac{dy}{dt} = v_y(y), & \text{or} & \frac{dy}{v_y(y)} = dt \\ \frac{dz}{dt} = v_z(z), & \text{or} & \frac{dz}{v_z(z)} = dt \end{array} \tag{15.36}$$

Since v_x depends only on x, v_y depends only on y and v_z depends only on z, the partial ODE's (15.36) can be integrated analytically between two arbitrary times t_n to t_{n+1}, viz.,

$$\begin{array}{c} \int_{x(t_n)}^{x(t_{n+1})} \frac{dx}{v_x(x)} = \int_{t_n}^{t_{n+1}} dt \\ \int_{y(t_n)}^{y(t_{n+1})} \frac{dy}{v_y(y)} = \int_{t_n}^{t_{n+1}} dt \\ \int_{z(t_n)}^{z(t_{n+1})} \frac{dz}{v_z(z)} = \int_{t_n}^{t_{n+1}} dt \end{array} \tag{15.37}$$

where $x(t_n)$, $y(t_n)$ and $z(t_n)$ are the particle coordinates at time t_n, and $x(t_{n+1})$, $y(t_{n+1})$ and $z(t_{n+1})$ are the particle coordinates at time t_{n+1}. Since A_x, A_y and A_z from (15.35) are constant for the difference cell, the integration (15.37) simply yields

$$\begin{array}{c} \ln \frac{v_x(x(t_{n+1}))}{v_x(x(t_n))} = A_x \Delta t_n \\ \ln \frac{v_y(y(t_{n+1}))}{v_y(y(t_n))} = A_y \Delta t_n \\ \ln \frac{v_z(z(t_{n+1}))}{v_z(z(t_n))} = A_z \Delta t_n \end{array} \tag{15.38}$$

where $\Delta t_n = t_{n+1} - t_n$ is the time increment. Assuming that the velocity at time t_n is known so that $v_x(t_n) = v_x(x(t_{n+1})) \neq 0$, $v_y(t_n) = v_y(y(t_{n+1})) \neq 0$ and $v_z(t_n) = v_z(z(t_{n+1})) \neq 0$, we can insert (15.35) into (15.38) to find explicit expressions for the particle location at the new time t_{n+1} as

$$\begin{aligned} x(t_{n+1}) &= x_1 + \tfrac{1}{A_x}[v_x(t_n)\exp(A_x \Delta t_n) - v_{x1}] \\ y(t_{n+1}) &= y_1 + \tfrac{1}{A_y}[v_y(t_n)\exp(A_y \Delta t_n) - v_{y1}] \\ z(t_{n+1}) &= z_1 + \tfrac{1}{A_z}[v_z(t_n)\exp(A_z \Delta t_n) - v_{z1}] \end{aligned} \tag{15.39}$$

provided that $A_x \neq 0$, $A_y \neq 0$ and $A_z \neq 0$. It is to be emphasized that the integration (15.38) is only valid when the values of A_x, A_y and A_z are constant and nonzero. As a consequence, the time increment Δt_n must be selected not larger than the particle crosses a cell surface. However, we can estimate directly from (15.38) the maximum time increment Δt_e for a cell which results when a particle entering the cell surface at the entry point (x_p, y_p, z_p) will reach an exit point (x_e, y_e, z_e) at a cell surface (Fig. 15.31). Assuming that the velocity components at the entry point p and time t_n are given by v_{xp}, v_{yp} and v_{zp}, we can determine from (15.38) potential time increments which are possible to reach any exit on the cell surfaces. To demonstrate the procedure we consider the situation shown in Fig. 15.31, where the entry point p is located on the front face of the 3D cell and five choices (in 2D three choices) for reaching an exit on cell surfaces are possible. Let Δt_y be the time increment to reach the opposite face, Δt_z^T and Δt_z^B be the increments to reach either the top or the bottom face, respectively, and Δt_x^L and Δt_x^R be the increments to reach either the left or the right face, respectively, we obtain

$$\begin{aligned} \Delta t_x^L &= \tfrac{1}{A_x}\ln\tfrac{v_{x1}}{v_{xp}}, \quad & \Delta t_x^R &= \tfrac{1}{A_x}\ln\tfrac{v_{x2}}{v_{xp}} \\ \Delta t_y &= \tfrac{1}{A_y}\ln\tfrac{v_{y2}}{v_{yp}} & & \\ \Delta t_z^T &= \tfrac{1}{A_z}\ln\tfrac{v_{z2}}{v_{zp}}, \quad & \Delta t_z^B &= \tfrac{1}{A_z}\ln\tfrac{v_{z1}}{v_{zp}} \end{aligned} \tag{15.40}$$

and the permissible time step is selected as the smallest increment:

$$\Delta t_n \leq \Delta t_e = \min(\Delta t_x^L, \Delta t_x^R, \Delta t_y, \Delta t_z^T, \Delta t_z^B) \tag{15.41}$$

Choosing $\Delta t_n = \Delta t_e$ in (15.39) we can immediately determine the coordinates of the exit point (x_e, y_e, z_e) as

$$\begin{aligned} x_e &= x_1 + \tfrac{1}{A_x}[v_{xp}\exp(A_x \Delta t_e) - v_{x1}] \\ y_e &= y_1 + \tfrac{1}{A_y}[v_{yp}\exp(A_y \Delta t_e) - v_{y1}] \\ z_e &= z_1 + \tfrac{1}{A_z}[v_{zp}\exp(A_z \Delta t_e) - v_{z1}] \end{aligned} \tag{15.42}$$

The above procedure is repeated for each cell where the exit point of one cell forms the entry point of the other adjacent cell until the particle leaves the computational domain or stops at internal boundaries. Heuristic rules are applied [425] to treat discontinuous velocity components at the interfaces between adjacent cells, e.g., the velocities at the interface could have opposite directions. Stagnation points and pumping well conditions require special attention. To overcome the restriction to steady-state flow, Pollock's algorithm approximates time-varying velocities by a stepwise updating in time, i.e., at each time step the velocity must be treated as a steady-state solution. Another seriously restrictive feature of Pollock's method is the requirement for decoupling the particle tracking ODE's via separation of variables as given in (15.36), which is only achievable for a simple linear velocity interpolation typical in regular grid-cell geometries. Contrarily, in FEM the element interpolation of velocities, even for linear functions, implies mixed interpolation terms and can involve nonlinear dependencies arising, e.g., from variable-density and variably saturated flow, the decoupling like (15.36) is not possible in general for unstructured meshes and more complex flow conditions. Nevertheless, for linear flow problems Pollock's method has shown to be extensible to irregular meshes [99, 222, 424] if the algorithm is performed in transformed coordinates based on a one-to-one mapping of Euclidean $\boldsymbol{x}-$space to the local $\boldsymbol{\eta}-$space similar to isoparametric transformations (cf. Sect. 8.8.1). Accordingly, the particle tracking ODE's are formulated in local coordinates $\boldsymbol{\eta}^T = (\xi \ \eta \ \zeta)$ for a finite element e as

$$\begin{pmatrix} \frac{d\xi}{dt} \\ \frac{d\eta}{dt} \\ \frac{d\zeta}{dt} \end{pmatrix} = (\boldsymbol{J}^e)^{-1} \cdot \begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \\ \frac{dz}{dt} \end{pmatrix} = (\boldsymbol{J}^e)^{-1} \cdot \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \tag{15.43}$$

where $(\boldsymbol{J}^e)^{-1}$ is the inverse element Jacobian defined in (8.119). To make analytical integration of (15.43) possible, the Jacobian has to be approximated by constants to get an analytically tractable set of uncoupled pathline ODE's expressed in their local coordinates:

$$\begin{aligned} \frac{d\xi}{dt} &= v_\xi(\xi) \\ \frac{d\eta}{dt} &= v_\eta(\eta) \\ \frac{d\zeta}{dt} &= v_\zeta(\zeta) \end{aligned} \tag{15.44}$$

But, these transformed formulations are still restricted to simple element geometries (exceptions of constant Jacobians are discussed in Appendix H) and do not seem to make Pollock's tracking method substantially more flexible and applicable to unstructured meshes. Furthermore, it cannot remedy its basic deficiency with respect to the restriction to purely linear velocity relations and (at least elementwise) steady-state flow conditions. Much more flexibility at comparable accuracy, however, is provided by numerical tracking approaches to be studied next, which are applicable with *any* scheme of velocity interpolation in space and time.

15.2.3 Numerical 4th-Order Runge-Kutta Tracking Method

Practical tracking needs solution of the governing ODE (15.30) for generally nonuniform and unsteady flow fields $\boldsymbol{v}(\boldsymbol{x}, t)$. Analytical and semianalytical solutions are limited to cases with simple geometry, (elementwise) steady-state flow conditions and strictly linear velocity dependency. Numerical tracking methods, in

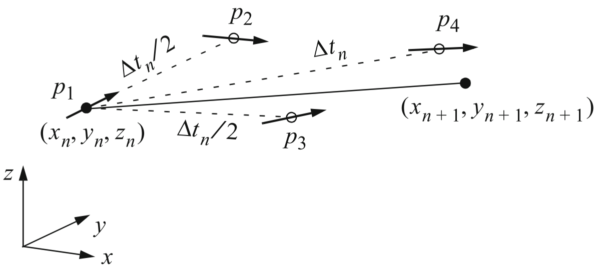


Fig. 15.32 Representation of the 4th-order Runge-Kutta method. In each time step, the velocity $\boldsymbol{v}(\boldsymbol{x}, t)$ is evaluated four times: once at initial point p_1, twice at trial midpoints (p_2 and p_3) and once at a trial endpoint p_4. Taking these velocities the final position for the next step (shown as a *filled dot*) is computed (Modified from [587])

particular higher-order Runge-Kutta methods, have shown very powerful, flexible and efficient [424]. Combined with adaptive stepping control they provide excellent accuracy needed for particle tracking computations in general flow fields with spatially and temporarily variable velocity. The numerical integration (15.31) can be performed either in global or local coordinate systems. In the following we describe the favorite 4th-order Runge-Kutta particle tracking integrator, which is explicit in time and basically operates in a global coordinate system (Cartesian coordinates in 3D and 2D as well as $r - z$–cylindrical coordinates of 2D meridional domain for axisymmetric problems).

The explicit 4th-order Runge-Kutta method is one of the most commonly used strategies for solving ODE's of type (15.30), see, e.g., [430] for more. It requires four evaluations of the velocity $\boldsymbol{v}(\boldsymbol{x}, t)$ for each particle tracking step Δt_n: once at the initial point p_1, twice at two trial midpoints (p_2 and p_3) and once at a trial endpoint p_4 as illustrated in Fig. 15.32. One Runge-Kutta step for the movement of a particle from the position $\boldsymbol{x}_n$ at given time plane n to the position $\boldsymbol{x}_{n+1}$ at the new time plane $n + 1$ with a time step increment $\Delta t_n = t_{n+1} - t_n$ then reads

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \tfrac{1}{6}(\overline{\boldsymbol{v}}_{p_1} + 2\overline{\boldsymbol{v}}_{p_2} + 2\overline{\boldsymbol{v}}_{p_3} + \overline{\boldsymbol{v}}_{p_4}) + \mathcal{O}(\Delta t_n^5) \tag{15.45}$$

where

$$\begin{aligned}
\overline{\boldsymbol{v}}_{p_1} &= \Delta t_n\, \boldsymbol{v}(\boldsymbol{x}_n, t_n) \\
\overline{\boldsymbol{v}}_{p_2} &= \Delta t_n\, \boldsymbol{v}(\boldsymbol{x}_n + \tfrac{\overline{\boldsymbol{v}}_{p1}}{2}, t_n + \tfrac{\Delta t_n}{2}) \\
\overline{\boldsymbol{v}}_{p_3} &= \Delta t_n\, \boldsymbol{v}(\boldsymbol{x}_n + \tfrac{\overline{\boldsymbol{v}}_{p2}}{2}, t_n + \tfrac{\Delta t_n}{2}) \\
\overline{\boldsymbol{v}}_{p_4} &= \Delta t_n\, \boldsymbol{v}(\boldsymbol{x}_n + \overline{\boldsymbol{v}}_{p_3}, t_n + \Delta t_n)
\end{aligned} \tag{15.46}$$

Since the accuracy of the integration method is primarily dependent on the tracking step size Δt_n, an adaptive time step size control is desirable to monitor the integration error of the particle tracking solution and achieve efficient computations. Regarding the Runge-Kutta method, the most straightforward techniques for

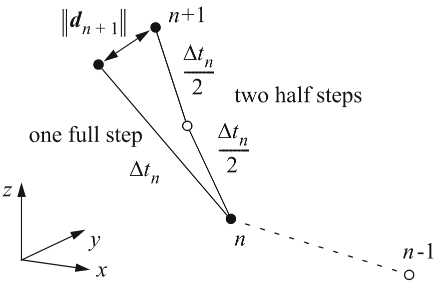


Fig. 15.33 Step doubling for adaptive step size control in the 4th-order Runge-Kutta method

adaptive time step size control is *step doubling* [430]. In such a step doubling procedure, a tracking step Δt_n is always taken twice: once as a full step and once as two half steps. Thus, the resulting difference in the particle location is found as $\boldsymbol{d}_{n+1} = \boldsymbol{x}^{\Delta}_{n+1} - \boldsymbol{x}^{\Delta/2}_{n+1}$, where $\boldsymbol{x}^{\Delta}_{n+1}$ is the position for the full step according to (15.45) and $\boldsymbol{x}^{\Delta/2}_{n+1}$ is the position after two half steps (Fig. 15.33). If Δt_n is small enough, $\|\boldsymbol{d}_{n+1}\|$ will also be small. Since the accuracy of the 4th-order Runge-Kutta method implies an error of $\mathcal{O}(\Delta t_n^5)$, the following error estimation holds

$$\left(\frac{\Delta t_n}{\Delta t_{n+1}}\right)^5 = \frac{\|\boldsymbol{d}_{n+1}\|}{\|\boldsymbol{d}_{n+2}\|} \tag{15.47}$$

where $\boldsymbol{d}_{n+2}$ is the difference in particle location associated with the forthcoming step length Δt_{n+1}. Based on the requirement that an error norm for the forthcoming step should be equal to a pre-set tolerance error $\epsilon_{\|} = \|\boldsymbol{d}_{n+2}\|$, e.g., 10^{-4}, we can utilize (15.47) to estimate the new time step size Δt_{n+1}. The following pragmatic relation results

$$\Delta t_{n+1} = \begin{cases} \varsigma_s \Delta t_n \left(\frac{\epsilon_{\|}}{\|\boldsymbol{d}_{n+1}\|}\right)^{1/5} & \text{if} \quad \|\boldsymbol{d}_{n+1}\| \le \epsilon_{\|} \\ \varsigma_s \Delta t_n \left(\frac{\epsilon_{\|}}{\|\boldsymbol{d}_{n+1}\|}\right)^{1/4} & \text{if} \quad \|\boldsymbol{d}_{n+1}\| > \epsilon_{\|} \end{cases} \tag{15.48}$$

where ς_s is a 'safety factor', usually set to 0.9. Note in (15.48), when the step size must be reduced at $\|\boldsymbol{d}_{n+1}\| > \epsilon_{\|}$, the new prediction is deemed only appropriate for an error $\mathcal{O}(\Delta t_n^4)$ so that the exponent becomes $1/4$ instead of $1/5$ [430]. In the practical computation the error norm $\|\boldsymbol{d}_{n+1}\|$ is taken in (15.48) as a maximum error norm defined as

$$\|\boldsymbol{d}_{n+1}\|_{L\infty} = \max_i \left(\frac{|x^{\Delta}_{i,n+1} - x^{\Delta/2}_{i,n+1}|}{|x_{i,n}| + |\Delta t_n v_{i,n}|}\right), \quad (i = 1, 2, 3) \tag{15.49}$$

where vectors $\boldsymbol{x}$ and $\boldsymbol{v}$ are written in index notation (Sect. 2.1.1).

In the context of FEM the 4th-order Runge-Kutta method is implemented with the following practical rules:

- The Runge-Kutta steps are performed within a finite element until the following criteria are satisfied:
  - If the predicted point $\boldsymbol{x}_{n+1}$ is located on the element boundary or outside the current element, the exit point of the element is computed (see further below). The pathline ends if the element boundary is part of a global (inner or outer) boundary Γ of the computational domain $\bar{\Omega}$, otherwise the adjacent element is determined and the Runge-Kutta stepping is proceeded, where the exit point is taken as entry point for the new element.
  - Tracking stops if an effective stepping velocity v_{eff} falls below a minimum velocity:

$$v_{\text{eff}} = \frac{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_n\|}{\Delta t_n} \le \tfrac{1}{100} v_{\min} \tag{15.50}$$

    where $v_{\min}$ is determined as the measure of the smallest velocity in the mesh. This typically occurs if a pathline approaches to a pumping well (nodal singularity). If the particle point is sufficiently close to a pumping well, the situation usually arises that the first half step in the advancing Runge-Kutta process will generate a particle point behind the well, while the second half step will locate the point more closely again to the well position due to the opposite velocity direction, which can lead to an oscillation around the well point. Criterion (15.50) is efficient to terminate pathlines at wells in avoiding a too costly oscillatory approach.
  - A maximum number of integration steps within an element is exceeded (default value: 200). This is beneficial to avoid oscillations in the tracking process for pumping wells surrounded by coarse elements.
  - For a steady-state velocity field the maximum travel time (default: 10^{30} d) is exceeded.
- The adaptive stepping of the Runge-Kutta method is exclusively monitored via the dimensionless tolerance error $\epsilon_{\|}$. It significantly affects the accuracy and the performance of the numerical integration and must be appropriately chosen. Values of $\epsilon_{\|} = 10^{-4}, \ldots, 10^{-6}$ have shown useful.
- The initial step size Δt_0 is assessed for an element e, where the tracking starts, by

$$\Delta t_0 = 0.3 \frac{h^e}{v^e_{\max}} \tag{15.51}$$

  where h^e is the characteristic element length (8.239) and $v^e_{\max}$ is the measure of the maximum velocity occurring in the element e. The empirical factor of 0.3 ensures that at least three integration steps are taken to pass through the element.
- During the Runge-Kutta stepping at each particle point $\boldsymbol{x}$ and time t the velocity $\boldsymbol{v}(\boldsymbol{x}, t)$ is obtained via interpolation by using three working steps (Sect. 15.4): (1) Find from the given global coordinate $\boldsymbol{x}$ the corresponding element e and determine the local coordinates $\boldsymbol{\eta}$, (2) determine the element shape functions $N^e_J(\boldsymbol{\eta})$ at given $\boldsymbol{\eta}$ and (3) interpolate the velocity $\boldsymbol{v}^e(\boldsymbol{x}(\boldsymbol{\eta}), t) = \sum_J N^e_J(\boldsymbol{\eta})\boldsymbol{v}^e_J(t)$ from the known nodal velocities $\boldsymbol{v}^e_J(t)$ at element nodes $J = 1, \ldots, N_{\text{BN}}$.
- To determine whether a point $\boldsymbol{x}$ is still located in the current element e, the associated local coordinates $\boldsymbol{\eta}$ are tested against their limits $+1/-1$ or $0/1$ in

dependence on the actual element type, extended by a small numerical tolerance of 10^{-7}, e.g., $-1-10^{-7} \leq \boldsymbol{\eta} \leq 1+10^{-7}$ applied to quadrilaterals (bricks).

- Suppose that $\boldsymbol{x}_{n+1}$ represents a predicted pathline point which is located outside the current element e and suppose that $\boldsymbol{x}_n$ represents the previous pathline point which is located within the element e, the exit point $\boldsymbol{x}_{n+1}^E$ is determined by clipping the particle line segment with the corresponding bounding lines (faces) of element e, such as

$$\boldsymbol{x}_{n+1}^E = \boldsymbol{x}_n + k(\boldsymbol{x}_{n+1} - \boldsymbol{x}_n) \tag{15.52}$$

  where the factor $0 \leq k \leq 1$ is determined from appropriate clipping algorithms. However, due to numerical noise it may happen that $\boldsymbol{x}_n$ is already located on the element face or slightly ahead so that k becomes negative. To handle this, the clipping is tested for $-10^{-6} \leq k \leq 1$.
- Travel times are marked by isochrons, which represent predefined time stages $t_p\,(p = 1,2,\ldots)$. Their coordinate $\boldsymbol{x}(t_p)$ results from linear interpolation between the previous coordinate $\boldsymbol{x}_n$ and predicted coordinate $\boldsymbol{x}_{n+1}$ of the pathline if $t_n \leq t_p \leq t_{n+1}$.
- The accuracy of the adaptive Runge-Kutta method can demand a large number of steps so that the generated number of pathline points considerably exceeds the resolution, which is sufficient for a graphical display of the pathline. A compression of the pathline points is commonly suited to reduce significantly the number of pathline points in accordance with the actual graphical resolution.

15.3 Streamline Computation

15.3.1 Introduction

The evaluation of velocity fields is of important interest in the finite-element flow analysis. Commonly, velocities $\boldsymbol{v}$ are derived as nodal quantities from primary variables by using suited smoothing techniques as described elsewhere, see Sects. 8.19.1, 9.7, 10.11, 11.7 and 14.4. If velocity $\boldsymbol{v}$ is known different methods are available to trace and visualize the flow field in a postprocessing procedure. The most general method concerns particle tracking (see preceding Sect. 15.2), where a pathline of an individual fluid particle is traced in space $\boldsymbol{x}$ and time t via a Lagrangian approach. Particle tracking methods are applicable both in 2D and 3D under very general flow conditions (presence of interior sinks/sources and/or BC's such as pumping wells and others). While they refer to individually moving particles which have to be appropriately assigned at starting positions, a continuous picture of the overall flow movement is sometimes difficult to attain, even if a large number of particles is traced.

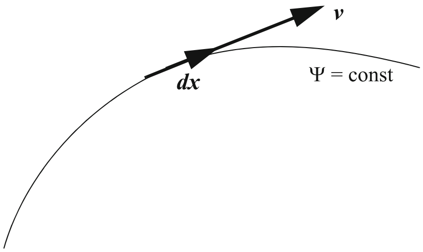


Fig. 15.34 Definition of a streamline

There are efficient, but specific alternative methods for limited cases in 2D applications. These methods represent *streamline* integrators, which facilitate the computation of flow pattern and distributed discharge through the flow systems in a direct way. The two most important streamline integrators will be described in the following.

15.3.2 Streamline and Streamfunction

The basic definitions of streamline and streamfunction as already introduced in Sect. 2.1.11 are briefly rediscussed in the present context. We follow the definitions of spatial variables and derivatives according to Sect. 2.1.6. A *streamline* is the locus of points that are everywhere tangent to the instantaneous velocity vector $\boldsymbol{v}$. This tangency requires that the cross product must give $\boldsymbol{v} \times d\boldsymbol{x} = \boldsymbol{0}$, where $d\boldsymbol{x}$ is a differential along a streamline (Fig. 15.34). Referring to 2D Cartesian coordinates with $\boldsymbol{v}^T = (v_1 \; v_2)$ and $d\boldsymbol{x}^T = (dx_1 \; dx_2)$, it yields

$$\frac{dx_1}{v_1} = \frac{dx_2}{v_2} \tag{15.53}$$

and similar for axisymmetric flow. Two streamlines cannot intersect, except where $\boldsymbol{v} = \boldsymbol{0}$. Since, by definition, no flow can cross a streamline it requires that the velocity vector field $\boldsymbol{v}$ have to be divergence-free (solenoidal), i.e.,

$$\nabla \cdot \boldsymbol{v} = 0 \tag{15.54}$$

That means the flow has to be steady-state and no distributed sources and sinks can exist in the flow domain. An equation that would describe such streamlines in a 2D and axisymmetric steady-state flow may be written in the form

$$\Psi = \begin{cases} \Psi(x_1, x_2) & \text{2D Cartesian} \\ \Psi(r, z) & \text{axisymmetric} \end{cases} \tag{15.55}$$

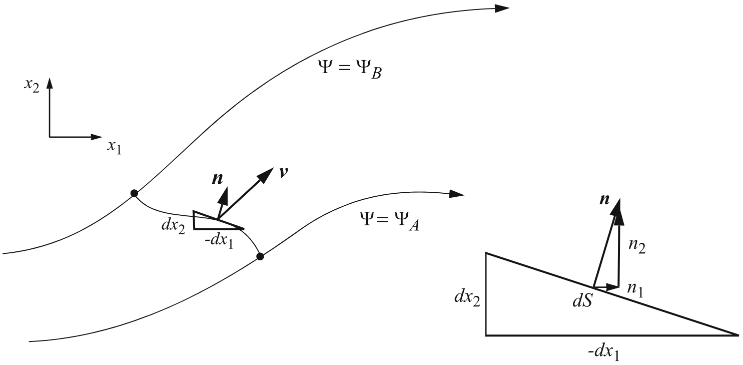


Fig. 15.35 Streamfunction in a plane flow

where Ψ is called the *streamfunction*. We note that a streamfunction analog doesn't exist for 3D flow. The following definition relates Ψ and the velocity components

$$\begin{aligned} v_1 &= \frac{\partial \Psi}{\partial x_2}, \quad v_2 = -\frac{\partial \Psi}{\partial x_1} \qquad \text{2D Cartesian} \\ v_r &= \frac{1}{r}\frac{\partial \Psi}{\partial z}, \; v_z = -\frac{1}{r}\frac{\partial \Psi}{\partial r} \qquad \text{axisymmetric} \end{aligned} \tag{15.56}$$

The definitions (15.56) automatically satisfy the condition of free divergence (15.54) in using (2.74). A major characteristic of the streamfunction is that the change $\delta\Psi$ in Ψ between two streamlines is equal to the volume flow rate between those streamlines. Let us consider two streamlines with values Ψ_A and Ψ_B as shown in Fig. 15.35, then the volume flow rate between the streamlines is (in 2D Cartesian)

$$\delta\Psi = \int_A^B \boldsymbol{v} \cdot \boldsymbol{n}\, dS = \int_A^B (v_1 n_1 + v_2 n_2)\, dS \tag{15.57}$$

where $\boldsymbol{n}^T = (n_1 \; n_2)$ is the unit normal vector, directed outward to the integration path dS. By geometry, we have the relations $n_1\, dS = dx_2$ and $n_2\, dS = -dx_1$, and (15.57) becomes

$$\delta\Psi = \int_A^B (v_1\, dx_2 - v_2\, dx_1) = \int_A^B d\Psi = \Psi_B - \Psi_A. \tag{15.58}$$

The flow rate between streamlines is the difference in their streamfunction values. This equation is also unaffected by the addition of an arbitrary constant to Ψ.

15.3.3 Streamline Integration via Vorticity Equation

For 2D and axisymmetric flows a very efficient approach of computing the streamfunction distribution for a given velocity field $\boldsymbol{v}$ is based on using the vorticity $\boldsymbol{\omega} = \nabla \times \boldsymbol{v}$, defined in (2.75). While in 3D $\boldsymbol{\omega}$ represents a general vector field, in 2D and for axisymmetric flows the following useful curl-relations hold

$$\omega = \|\boldsymbol{\omega}\| = \begin{cases} \dfrac{\partial v_2}{\partial x_1} - \dfrac{\partial v_1}{\partial x_2} & \text{2D } (x_1, x_2) \text{ Cartesian} \\ \dfrac{\partial v_r}{\partial z} - \dfrac{\partial v_z}{\partial r} & \text{axisymmetric } (r, z) \end{cases} \tag{15.59}$$

where ω is the (scalar) *vorticity function*. By substituting the streamfunction definition (15.56) into the vorticity equation (15.59) the following elliptic partial differential (Poisson-type) equation is obtained

$$-\nabla^2 \Psi = \begin{cases} \dfrac{\partial v_2}{\partial x_1} - \dfrac{\partial v_1}{\partial x_2} & \text{2D } (x_1, x_2) \text{ Cartesian} \\ \left(\dfrac{\partial v_z}{\partial r} - \dfrac{\partial v_r}{\partial z} \right) r & \text{axisymmetric } (r, z) \end{cases} \tag{15.60}$$

Equation (15.60) can be easily solved via FEM if formulating a unique boundary value problem of the domain Ω enclosed by the boundary Γ. Introducing finite element interpolation functions for the streamfunction and velocity components (exemplified for 2D Cartesian)

$$\left. \begin{aligned} \Psi &= \textstyle\sum_j N_j(\boldsymbol{x}) \Psi_j \\ v_1 &= \textstyle\sum_j N_j(\boldsymbol{x}) U_j \\ v_2 &= \textstyle\sum_j N_j(\boldsymbol{x}) V_j \end{aligned} \right\} \quad (j = 1, \ldots, N_{\mathrm{P}}) \tag{15.61}$$

where N_j represent the $j-$nodal finite element basis functions, Ψ_j, U_j and V_j correspond to the nodal values of the streamfunction and v_1, v_2-velocity components, respectively, the GFEM formulation of (15.60) leads to the following matrix system

$$\boldsymbol{A} \cdot \boldsymbol{\Psi} = \boldsymbol{B}(\boldsymbol{U}, \boldsymbol{V}) \tag{15.62}$$

to be solved for the nodal streamfunction vector $\boldsymbol{\Psi}^T = (\Psi_1 \ \Psi_2 \ \ldots \ \Psi_{N_{\mathrm{P}}})$ with known velocity vectors $\boldsymbol{U}^T = (U_1 \ U_2 \ \ldots \ U_{N_{\mathrm{P}}})$ and $\boldsymbol{V}^T = (V_1 \ V_2 \ \ldots \ V_{N_{\mathrm{P}}})$ on the RHS, in which

$$\begin{aligned}\boldsymbol{A} = A_{ij} &= \sum_e \int_{\Omega^e} \nabla N_i \cdot \nabla N_j \, d\Omega^e \\ \boldsymbol{B} = B_i \ &= \sum_e \Big(\int_{\Omega^e} N_i \big(\sum_j \tfrac{\partial N_j}{\partial x_1} V_j - \sum_j \tfrac{\partial N_j}{\partial x_2} U_j \big) d\Omega^e + \int_{\Gamma_N^e} N_i (\nabla \Psi \cdot \boldsymbol{n}) d\Gamma^e \Big)\end{aligned} \tag{15.63}$$

Similar expressions result for axisymmetric flow. The boundary integral in (15.63) vanishes because the flux normal to the streamline direction is zero, $\nabla\Psi \cdot \boldsymbol{n} = \frac{\partial \Psi}{\partial x_1} n_1 + \frac{\partial \Psi}{\partial x_2} n_2 = -v_2 n_1 + v_1 n_2 = 0$ if the velocity vector field $\boldsymbol{v}$ is divergence-free (15.54).

The matrix $\boldsymbol{A}$ is symmetric and sparse. The linear equation system (15.62) is easily solved by using standard matrix solvers. However, a suitable BC for Ψ is required. Practically, at only one node on the outer boundary Γ the streamfunction is set to a Dirichlet-type reference value of zero. The solution of (15.62) is restricted to a solenoidal 2D (or axisymmetric) velocity vector field $\boldsymbol{v}$, i.e., steady-state flow, no interior BC's (e.g., fluxes, wells) and absence of sinks and sources.

15.3.4 Streamline Integration via Boundary Integral

The streamline integration method is based on the numerically solution of differential (15.57) written in the form

$$\delta\Psi = \begin{cases} \int_A^B (v_1 n_1 + v_2 n_2) \, dS & \text{2D } (x_1, x_2) \text{ Cartesian} \\ 2\pi \int_A^B (v_r n_r + v_z n_z) \, r \, dS & \text{axisymmetric } (r, z) \end{cases} \tag{15.64}$$

where $\delta\Psi$ is the change in the streamfunction which is to be solved along a defined boundary. In the preferred method the computation of $\delta\Psi$ is carried out using (15.64) along each boundary of finite elements Γ^e, where the integration path $\widehat{AB}$ is taken as element edge. We consider a typical element boundary as shown in Fig. 15.36. The following finite element interpolations for element e are introduced (exemplified for 2D Cartesian)

$$\left.\begin{aligned} v_1^e &= \textstyle\sum_J N_J^e(\eta) U_J^e, \quad & x_1^e &= \textstyle\sum_J N_j^e(\eta) X_J^e \\ v_2^e &= \textstyle\sum_J N_J^e(\eta) V_J^e, \quad & x_2^e &= \textstyle\sum_J N_j^e(\eta) Y_J^e \end{aligned}\right\} \quad (J = 1, \ldots, N_{\text{BN}}) \tag{15.65}$$

where N_J^e are finite element shape functions and $U_J^e, V_J^e, X_J^e, Y_J^e$ are nodal point velocities and coordinates, respectively, associated with element e. In the finite element standard procedure the global coordinates (x_1, x_2) in 2D are transformed to local coordinates (ξ, η) (Fig. 15.36). For an infinitesimal line element dS it results (in 2D Cartesian)

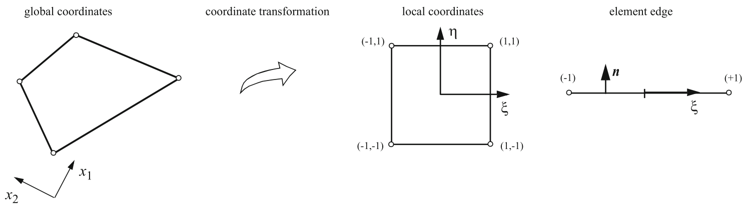


Fig. 15.36 Definition of element boundary for streamfunction computation exemplified for a quadrilateral element

$$dS = \sqrt{\left(\frac{\partial x_1}{\partial \xi}\right)^2 + \left(\frac{\partial x_2}{\partial \xi}\right)^2}\, d\xi = L\, d\xi \tag{15.66}$$

written for the local coordinate ($-1 \leq \xi \leq +1$) at element boundaries where $\eta = \pm 1$. In (15.66) L corresponds to the length of the boundary segment. Using (15.66) the unit normal vector can be expressed by

$$\boldsymbol{n} = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial x_2}{\partial S} \\ -\frac{\partial x_1}{\partial S} \end{pmatrix} = \frac{1}{L} \begin{pmatrix} \frac{\partial x_2}{\partial \xi} \\ -\frac{\partial x_1}{\partial \xi} \end{pmatrix} \tag{15.67}$$

Combining (15.66) and (15.67) with (15.64) the streamline integral along any finite element boundary of element e takes the form (exemplified for 2D Cartesian)

$$\delta \Psi^e = \int_{-1}^{+1} \left(\sum_J \frac{\partial N_J^e}{\partial \xi} Y_J^e \sum_J N_J^e U_J^e - \sum_J \frac{\partial N_J^e}{\partial \xi} X_J^e \sum_J N_J^e V_J^e \right) d\xi \tag{15.68}$$

where the element interpolation functions (15.65) are applied.

The change in the streamfunction along any element boundary is computed from (15.68) with known nodal velocity vectors $\boldsymbol{U}$, $\boldsymbol{V}$. The computation of the streamfunction for an entire finite element mesh is generated by applying (15.68) along successive element boundaries, starting at a node for which a reference value of Ψ has been specified. Unlike the vorticity integration method (Sect. 15.3.3) the present boundary integral method is only an element-by-element procedure, which is computationally efficient and does not require the solution of a matrix problem. However, the boundary integrator is also limited to solenoidal velocity fields, i.e., steady-state flow, no interior BC's (e.g., fluxes, wells) and absence of sinks and sources.

15.3.5 Discussion

While both streamline integrators are only limited to steady-state 2D (or axisymmetric) flow problems, where neither interior BC's (such as fluxes or pumping wells) nor sinks and sources should exist, they have shown very useful to visualize flow patterns in relation to the actual discharge distribution within a flow domain. Typical applications have been illustrated in Figs. 10.33, 10.38,10.43,11.5,11.9,11.15,11.21,11.24,11.28,14.13,15.27 and 15.28 of the preceding chapters. While the boundary integral method does not require the solution of a matrix problem, the vorticity equation integrator produces often more accurate results and has been proven more robust. Of particular relevance are streamline integrators in application to variable-density flow situations, where complex recirculating flow patterns (eddies) can occur, which cannot be easily detected and suitably visualized by using particle tracking methods. In such cases, although the density-coupled mass (or heat) transport process is transient, the flow field may be divergence-free at each time step (FEFLOW runs in the so-called *steady flow-transient transport* simulation mode) because the absence of storage (by fluid and skeleton compression) in the flow problem without interior flux BC's and sinks/sources.

15.4 Interpolation on Finite Elements

15.4.1 Scope

Computational results of a finite element solution are given in form of primary variables (e.g., hydraulic head, species concentration, temperature) and secondary variables (e.g., velocity, saturation) at nodal points of a finite element mesh. In various postprocessing evaluations it is required to determine the variables at arbitrary location $\boldsymbol{x}_p \in \bar{\Omega}$ of a point p within the discretized domain $\bar{\Omega}$ by using appropriate interpolation on finite elements, where the following three working steps are necessitated (Fig. 15.37):

1. Find the element e within which the point p lies.
2. For the given global coordinates $\boldsymbol{x}_p$ of the point p, determine the corresponding local coordinate $\boldsymbol{\eta}_p$ of the point p in the element e.
3. Take the local coordinates $\boldsymbol{\eta}_p$ to interpolate a variable $\phi(\boldsymbol{x}, t)$ at point p on element basis, viz.,

$$\phi(\boldsymbol{x}_p(\boldsymbol{\eta}_p), t) = \sum_{J=1}^{N_{\mathrm{BN}}} N_J^e(\boldsymbol{\eta}_p)\, \phi_J^e(t) \tag{15.69}$$

   where N_J^e are the element shape functions at local node J and $\phi_J^e(t)$ are the nodal values of the variable at given time t.

While the first step can be easily performed by using standard selection techniques (e.g., bounding box and point-in-polygon search), the second step needs specific attention because a direct reversion of the basis functions is not possible for the

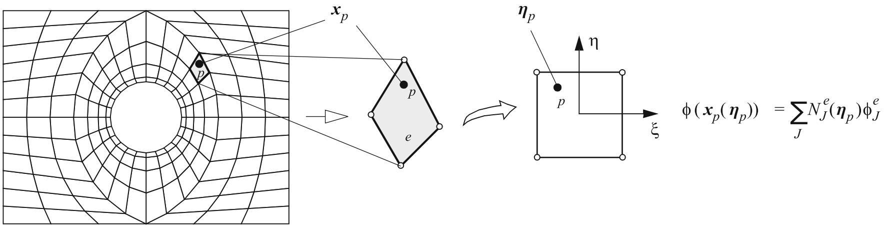

$\boldsymbol{x}_p$
p
p
e
$\boldsymbol{\eta}_p$
η
p
ξ
$\phi(\boldsymbol{x}_p(\boldsymbol{\eta}_p)) = \sum_J N_J^e(\boldsymbol{\eta}_p)\phi_J^e$

Fig. 15.37 Evaluation of a variable ϕ at a specified point p on a finite element mesh

most finite element types and a more general strategy is required to determine $\boldsymbol{\eta}_p$ from given $\boldsymbol{x}_p$ for any element type in an efficient way as discussed next. Once $\boldsymbol{\eta}_p$ is known the third step becomes simple. The local coordinates $\boldsymbol{\eta}_p$ are easily used, together with the shape functions, to evaluate the variable at the point p via standard finite element interpolation.

15.4.2 Accomplishment of Reverse Transformation for Isoparametric Elements

Global coordinates $\boldsymbol{x}$ are related to the local coordinates $\boldsymbol{\eta}$ of an isoparametric finite element e by mapping (8.71)

$$\boldsymbol{x}^e = \sum_{J=1}^{N_{\mathrm{BN}}} N_J^e(\boldsymbol{\eta})\,\boldsymbol{x}_J^e \tag{15.70}$$

exemplified in 3D Cartesian

$$\left.\begin{aligned} x^e &= \textstyle\sum_J N_J^e(\xi,\eta,\zeta)\,x_J^e \\ y^e &= \textstyle\sum_J N_J^e(\xi,\eta,\zeta)\,y_J^e \\ z^e &= \textstyle\sum_J N_J^e(\xi,\eta,\zeta)\,z_J^e \end{aligned}\right\} \quad (J = 1,\ldots,N_{\mathrm{BN}}) \tag{15.71}$$

where the shape functions $N_J^e(\boldsymbol{\eta})$ are listed for different isoparametric element types in Appendix G. The task is now to determine $\boldsymbol{\eta}_p^T = (\xi_p \;\; \eta_p \;\; \zeta_p)$ at a given point p with its global coordinates $\boldsymbol{x}_p^T = (x_p \;\; y_p \;\; z_p) = (x_p^e \;\; y_p^e \;\; z_p^e)$. It requires the reverse transformation, in which the Eqs. (15.71) must be solved for ξ_p, η_p and ζ_p simultaneously at point p. As seen from Tables G.2–G.4 of Appendix G, the local coordinates in the shape functions of 2D and 3D isoparametric elements are usually nonlinearly combined so that a direct reversion of (15.70) fails, except for linear triangular and linear tetrahedral elements for which the local coordinates could be reversely solved via simple resubstitution. A general and efficient solution is preferred here, which is applicable to all element types and based on the Newton iteration method (cf. Sect. 8.18.2).

The Newton iteration method applied to the 3D Cartesian (15.71) requires the solution of the three nonlinearly coupled equations, viz.,

$$\left.\begin{aligned} F_1(\boldsymbol{\eta}) &= \textstyle\sum_J N_J^e(\boldsymbol{\eta})\, x_J^e - x^e &= 0 \\ F_2(\boldsymbol{\eta}) &= \textstyle\sum_J N_J^e(\boldsymbol{\eta})\, y_J^e - y^e &= 0 \\ F_3(\boldsymbol{\eta}) &= \textstyle\sum_J N_J^e(\boldsymbol{\eta})\, z_J^e - z^e &= 0 \end{aligned}\right\} \quad (J = 1, \ldots, N_{\mathrm{BN}}) \tag{15.72}$$

for the local coordinates $\boldsymbol{\eta}^T = (\xi \;\; \eta \;\; \zeta)$ at known global coordinates (x^e, y^e, z^e) of given point p (for convenience we omit the $p-$subscript). Expanding the functions F_1, F_2 and F_3 as truncated Taylor series, the following Newton iteration scheme results

$$\begin{aligned} F_1(\boldsymbol{\eta}^{\tau+1}) &= F_1(\boldsymbol{\eta}^\tau) + \tfrac{\partial F_1(\boldsymbol{\eta}^\tau)}{\partial \xi^\tau}(\xi^{\tau+1} - \xi^\tau) + \tfrac{\partial F_1(\boldsymbol{\eta}^\tau)}{\partial \eta^\tau}(\eta^{\tau+1} - \eta^\tau) \\ &\quad + \tfrac{\partial F_1(\boldsymbol{\eta}^\tau)}{\partial \zeta^\tau}(\zeta^{\tau+1} - \zeta^\tau) - x^e = 0 \\ F_2(\boldsymbol{\eta}^{\tau+1}) &= F_2(\boldsymbol{\eta}^\tau) + \tfrac{\partial F_2(\boldsymbol{\eta}^\tau)}{\partial \xi^\tau}(\xi^{\tau+1} - \xi^\tau) + \tfrac{\partial F_2(\boldsymbol{\eta}^\tau)}{\partial \eta^\tau}(\eta^{\tau+1} - \eta^\tau) \\ &\quad + \tfrac{\partial F_2(\boldsymbol{\eta}^\tau)}{\partial \zeta^\tau}(\zeta^{\tau+1} - \zeta^\tau) - y^e = 0 \\ F_3(\boldsymbol{\eta}^{\tau+1}) &= F_3(\boldsymbol{\eta}^\tau) + \tfrac{\partial F_3(\boldsymbol{\eta}^\tau)}{\partial \xi^\tau}(\xi^{\tau+1} - \xi^\tau) + \tfrac{\partial F_3(\boldsymbol{\eta}^\tau)}{\partial \eta^\tau}(\eta^{\tau+1} - \eta^\tau) \\ &\quad + \tfrac{\partial F_3(\boldsymbol{\eta}^\tau)}{\partial \zeta^\tau}(\zeta^{\tau+1} - \zeta^\tau) - z^e = 0 \end{aligned} \tag{15.73}$$

or written in matrix form

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \cdot \begin{pmatrix} \Delta\xi^\tau \\ \Delta\eta^\tau \\ \Delta\eta^\tau \end{pmatrix} = \begin{pmatrix} a_{14} \\ a_{24} \\ a_{34} \end{pmatrix} \tag{15.74}$$

where τ is the iteration counter, $a_{11} = \partial F_1(\boldsymbol{\eta}^\tau)/\partial \xi^\tau$, $\Delta\xi^\tau = \xi^{\tau+1} - \xi^\tau$, $a_{12} = \partial F_1(\boldsymbol{\eta}^\tau)/\partial \eta^\tau$, $\Delta\eta^\tau = \eta^{\tau+1} - \eta^\tau$, $a_{13} = \partial F_1(\boldsymbol{\eta}^\tau)/\partial \zeta^\tau$, $\Delta\zeta^\tau = \zeta^{\tau+1} - \zeta^\tau$, $a_{14} = x^e - F_1(\boldsymbol{\eta}^\tau)$ and so forth. Using Cramer's rule we solve

$$\begin{aligned} \Delta\xi^\tau &= \frac{1}{|a|}\big[a_{14}(a_{22}a_{33} - a_{32}a_{23}) - a_{24}(a_{12}a_{33} - a_{32}a_{13}) + a_{34}(a_{12}a_{23} - a_{22}a_{13})\big] \\ \Delta\eta^\tau &= \frac{1}{|a|}\big[-a_{14}(a_{21}a_{33} - a_{31}a_{23}) + a_{24}(a_{11}a_{33} - a_{31}a_{13}) - a_{34}(a_{11}a_{23} - a_{21}a_{13})\big] \\ \Delta\zeta^\tau &= \frac{1}{|a|}\big[a_{14}(a_{21}a_{32} - a_{31}a_{22}) - a_{24}(a_{11}a_{32} - a_{31}a_{12}) + a_{34}(a_{11}a_{22} - a_{21}a_{12})\big] \\ &\text{with} \\ |a| &= a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{21}(a_{12}a_{33} - a_{32}a_{13}) + a_{31}(a_{12}a_{23} - a_{22}a_{13}) \neq 0 \end{aligned} \tag{15.75}$$

for 3D elements.[4] The partial derivative terms are evaluated numerically by perturbation of the variables, e.g.,

[4] Equivalently, for 2D finite elements the Newton iteration scheme results

$$\frac{\partial F_1(\boldsymbol{\eta}^\tau)}{\partial \xi^\tau} \approx \frac{F_1(\xi^\tau + \delta, \eta^\tau, \zeta^\tau) - F_1(\xi^\tau, \eta^\tau, \zeta^\tau)}{\delta} \tag{15.76}$$

and similar for the remaining derivatives appearing in (15.73), where the increment δ has shown sufficient with 0.01.

To accelerate the numerical evaluations of the functions $F_1(\boldsymbol{\eta}^\tau)$, $F_2(\boldsymbol{\eta}^\tau)$ and $F_3(\boldsymbol{\eta}^\tau)$, they are suitably separated into constant and variable parts. We replace (15.72) by

$$\left.\begin{aligned} F_1(\boldsymbol{\eta}) &= \textstyle\sum_J C_{1J}^e(x_J^e)\, f_J^e(\boldsymbol{\eta}) - x^e &= 0 \\ F_2(\boldsymbol{\eta}) &= \textstyle\sum_J C_{2J}^e(y_J^e)\, f_J^e(\boldsymbol{\eta}) - y^e &= 0 \\ F_3(\boldsymbol{\eta}) &= \textstyle\sum_J C_{3J}^e(z_J^e)\, f_J^e(\boldsymbol{\eta}) - z^e &= 0 \end{aligned}\right\} \quad (J = 1, \ldots, N_{\mathrm{BN}}) \tag{15.77}$$

or written in a compact form as

$$\boldsymbol{F}(\boldsymbol{\eta}) = \boldsymbol{C}^e(\boldsymbol{x}^e) \cdot \boldsymbol{f}^e(\boldsymbol{\eta}) - \boldsymbol{x}^e = \boldsymbol{0} \tag{15.78}$$

where the nodal N_{BN} vector $\boldsymbol{f}^e(\boldsymbol{\eta})$ covers the variable factors and the nodal $D \times N_{\mathrm{BN}}$ matrix $\boldsymbol{C}^e(\boldsymbol{x}^e)$ contains the constant factors for an element e. The latter is expressed as

$$\boldsymbol{C}^e(\boldsymbol{x}^e) = \boldsymbol{X}^e(\boldsymbol{x}^e) \cdot \boldsymbol{Q}^e \tag{15.79}$$

where $\boldsymbol{X}^e(\boldsymbol{x}^e)$ represents the $D \times N_{\mathrm{BN}}$ matrix of constant coordinate entities of element e given for 3D Cartesian ($D = 3$) as

$$\boldsymbol{X}^e(\boldsymbol{x}^e) = \begin{pmatrix} x_1 & x_2 & \ldots & x_{N_{\mathrm{BN}}} \\ y_1 & y_2 & \ldots & y_{N_{\mathrm{BN}}} \\ z_1 & z_2 & \ldots & z_{N_{\mathrm{BN}}} \end{pmatrix} \tag{15.80}$$

$$\begin{aligned} F_1(\boldsymbol{\eta}^{\tau+1}) &= F_1(\boldsymbol{\eta}^\tau) + \tfrac{\partial F_1(\boldsymbol{\eta}^\tau)}{\partial \xi^\tau}(\xi^{\tau+1} - \xi^\tau) + \tfrac{\partial F_1(\boldsymbol{\eta}^\tau)}{\partial \eta^\tau}(\eta^{\tau+1} - \eta^\tau) - x^e = 0 \\ F_2(\boldsymbol{\eta}^{\tau+1}) &= F_2(\boldsymbol{\eta}^\tau) + \tfrac{\partial F_2(\boldsymbol{\eta}^\tau)}{\partial \xi^\tau}(\xi^{\tau+1} - \xi^\tau) + \tfrac{\partial F_2(\boldsymbol{\eta}^\tau)}{\partial \eta^\tau}(\eta^{\tau+1} - \eta^\tau) - y^e = 0 \end{aligned}$$

or

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} \Delta\xi^\tau \\ \Delta\eta^\tau \end{pmatrix} = \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix}$$

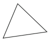
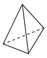
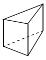
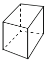
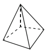
and

$$\begin{aligned} \Delta\xi^\tau &= \frac{1}{|a|}(a_{13}a_{22} - a_{23}a_{12}) \\ \Delta\eta^\tau &= \frac{1}{|a|}(a_{11}a_{23} - a_{21}a_{13}) \end{aligned}$$

with

$$|a| = a_{11}a_{22} - a_{21}a_{12} \neq 0$$

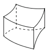
Table 15.2 Constant nodal matrix $\boldsymbol{Q}^e$ and variable nodal vector $\boldsymbol{f}^e(\boldsymbol{\eta})$ for different element types

Type	N_{BN}	$\boldsymbol{Q}^e$	$\boldsymbol{f}^e(\boldsymbol{\eta})$
	3	$\begin{pmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \end{pmatrix}^e$
	4	$\frac{1}{4}\begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \\ \xi \\ \eta \end{pmatrix}^e$
	8	$\frac{1}{4}\begin{pmatrix} -1 & 0 & 0 & 1 & 1 & 1 & -1 & -1 \\ 2 & 0 & -2 & 2 & 0 & 0 & 2 & 0 \\ -1 & 0 & 0 & 1 & 1 & -1 & -1 & 1 \\ 2 & 2 & 0 & 0 & 2 & 0 & 0 & -2 \\ -1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 2 & 0 & 2 & 2 & 0 & 0 & -2 & 0 \\ -1 & 0 & 0 & 1 & 1 & -1 & 1 & -1 \\ 2 & -2 & 0 & 0 & 2 & 0 & 0 & 2 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \\ \xi^2 \\ \eta^2 \\ \xi\eta \\ \xi^2\eta \\ \xi\eta^2 \end{pmatrix}^e$
	4	$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \\ \zeta \end{pmatrix}^e$
	6	$\frac{1}{2}\begin{pmatrix} 1 & -1 & -1 & 1 & -1 & -1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \\ \zeta \\ \xi\zeta \\ \eta\zeta \end{pmatrix}^e$
	8	$\frac{1}{8}\begin{pmatrix} 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \\ \zeta \\ \xi\eta \\ \xi\zeta \\ \eta\zeta \\ \xi\eta\zeta \end{pmatrix}^e$
	5	$\frac{1}{4}\begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 4 & 0 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \\ \xi\eta \\ \zeta \\ \frac{\xi\eta\zeta}{1-\zeta} \end{pmatrix}^e$

(continued)

and the $N_{\mathrm{BN}} \times N_{\mathrm{BN}}$ matrix $\boldsymbol{Q}^e$ contains constant element factors. The matrix $\boldsymbol{Q}^e$ and vector $\boldsymbol{f}^e(\boldsymbol{\eta})$ are summarized for different 2D and 3D finite element types in Table 15.2. Note that the linear 5-node pyramidal element requires one extra variable due to the nonlinear dependency in the shape functions.

Table 15.2 (continued)

Type	N_{BN}	$\boldsymbol{Q}^e$	$\boldsymbol{f}^e(\boldsymbol{\eta})$
	20	$\frac{1}{8}\begin{pmatrix} -2 & 1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 0 & 0 & 0 \\ 2 & 0 & -2 & 2 & -2 & 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & -2 & 0 & 0 \\ -2 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 0 & 0 & 0 \\ 2 & 2 & 0 & 2 & 0 & 0 & 0 & -2 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 2 & 0 \\ -2 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 2 & 2 & -2 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 \\ -2 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 0 & 0 & 0 \\ 2 & -2 & 0 & 2 & 0 & 0 & 0 & -2 & 2 & -2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & -2 & 0 \\ 2 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 2 & 2 & 0 & 0 & 0 & -2 & 0 & 0 & 2 \\ 2 & 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & -2 \\ 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & 0 & 0 & 0 & -2 & 0 & 0 & 2 \\ 2 & -2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 2 & -2 & 0 & 0 & 0 & 2 & 0 & 0 & -2 \\ -2 & 1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & -2 & -2 & -2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 \\ -2 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 0 & 0 & 0 \\ 2 & 2 & 0 & -2 & 0 & 0 & 0 & -2 & -2 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & -2 & 0 \\ -2 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 2 & -2 & -2 & -2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & -2 & 0 & 0 \\ -2 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 0 & 0 & 0 \\ 2 & -2 & 0 & -2 & 0 & 0 & 0 & -2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 2 & 0 \end{pmatrix}^e$	$\begin{pmatrix} 1 \\ \xi \\ \eta \\ \zeta \\ \xi^2 \\ \xi^2\eta \\ \xi^2\zeta \\ \eta^2 \\ \xi\eta^2 \\ \eta^2\zeta \\ \zeta^2 \\ \xi\zeta^2 \\ \eta\zeta^2 \\ \xi\eta\zeta \\ \xi^2\eta\zeta \\ \xi\eta^2\zeta \\ \xi\eta\zeta^2 \\ \eta\zeta \\ \xi\zeta \\ \xi\eta \end{pmatrix}^e$

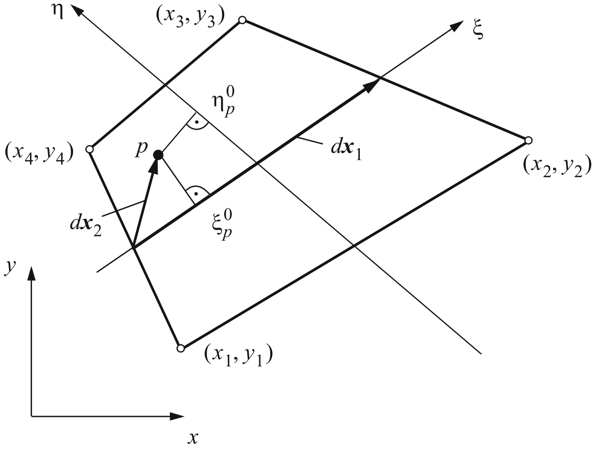


Fig. 15.38 Projecting point p onto element base lines

To utilize the rapid quadratic convergence of the Newton iteration, a good initial guess of the local coordinates $\boldsymbol{\eta}^{\tau=0}$ for the given global coordinate $\boldsymbol{x}_p^e$ of point p is necessary. A good initialization is attainable by point projection onto the base lines for planar or prismatic finite elements as shown in Fig. 15.38. Let $d\boldsymbol{x}_1^T = (dx_1 \;\; dy_1)$ be the base line vector and $d\boldsymbol{x}_2^T = (dx_2 \;\; dy_2)$ be the vector spanning between the point p and the first point of the base line, we obtain by simple vector projection (2.25)

$$\xi_p^0 = 2\frac{dx_1 dx_2 + dy_1 dy_2}{dx_1^2 + dy_1^2} - 1 \tag{15.81}$$

and similarly for η_p^0. For prismatic 3D elements the ζ_p^0 coordinate can be estimated by interpolation between top and bottom elevation z^T and z^B, respectively, of a vertical baseline such as

$$\zeta_p^0 = \frac{2z_p^e - z^T - z^B}{z^T - z^B} \tag{15.82}$$

This initialization has been found to be effective and the Newton method rapidly converges usually within ten iterations providing up to seven decimal places of accuracy.

Appendix A
Nomenclature

A.1 Roman Letters

A	L^2	area;
A^α	L^2T^{-2}	Helmholtz free energy of $\alpha-$phase;
B	L	aquifer thickness;
b	L	hydraulic aperture;
$b_k^\dagger$	$(ML^{-3})^{1-b_k^\ddagger}$	Freundlich sorption coefficient of species k;
$b_k^\ddagger$	1	Freundlich sorption exponent of species k;
b_k^p		rate constants of species k, $(p = 0, 1, \dots, N)$;
C	$L^{1/2}T^{-1}$	Chezy roughness coefficient;
C	L^{-1}	moisture capacity;
C^{-1}	L	inverse moisture capacity;
C_k	ML^{-3}	mass concentration of species k;
C_{ks}	ML^{-3}	maximum mass concentration of species k;
C_{kw}	ML^{-3}	prescribed concentration of species k at well point $\boldsymbol{x}_w$;
Cr	1	Courant number;
c	$L^2T^{-2}\Theta^{-1}$	specific heat capacity;
c_F	1	Forchheimer form-drag constant;
D		spatial dimension;
D	L	diameter or thickness;
$\mathcal{D}$		2nd-order differential operator or dispersion coefficient;
$\boldsymbol{D}_k$	L^2T^{-1}	tensor of hydrodynamic dispersion of species k;
$\boldsymbol{D}_k^\star$	L^2T^{-1}	nonlinear (extended) tensor of hydrodynamic dispersion of species k;
$\bar{\boldsymbol{D}}_k$	L^3T^{-1}	$= B\,\boldsymbol{D}_k$, depth-integrated tensor of hydrodynamic dispersion of species k;

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

$\boldsymbol{D}_{k,0}$	L^2T^{-1}	tensor of diffusion of species k;
$\boldsymbol{D}_{\text{mech}}$	L^2T^{-1}	tensor of mechanical dispersion;
$\bar{\boldsymbol{D}}_{\text{mech}}$	L^3T^{-1}	$= B\,\boldsymbol{D}_{\text{mech}}$, depth-integrated tensor of mechanical dispersion;
D_k	L^2T^{-1}	coefficient of molecular diffusion of species k in porous medium;
$\breve{D}_k$	L^2T^{-1}	coefficient of molecular diffusion of species k in open fluid body;
D_λ	L^2T^{-1}	$= [\varepsilon\Lambda + (1-\varepsilon)\Lambda^s]/(\varepsilon\rho c)$, thermal diffusivity;
$\boldsymbol{d}$	T^{-1}	$= \frac{1}{2}[\nabla\boldsymbol{v} + (\nabla\boldsymbol{v})^T]$, rate of deformation (strain) tensor;
d	L	characteristic length, thickness or diameter;
da	L^2	microscopic differential area;
dS	L^2	projected planar area of averaging volume;
dV	L^3	averaging volume;
dv	L^3	microscopic differential volume;
$\mathcal{E}$	ML^2T^{-2}	internal (thermal) energy;
E	L^2T^{-2}	specific internal (thermal) energy density;
E	$ML^{-1}T^{-2}$	Young's (or elastic) modulus of the solid phase;
$E^{\#}$	L^2T^{-2}	activation energy;
$\boldsymbol{e}$	1	$= -\boldsymbol{g}/\Vert\boldsymbol{g}\Vert$, gravitational unit vector;
$\boldsymbol{e}_{(\xi,\eta,\zeta)}$	1	$\boldsymbol{e}-$vector with respect to local coordinates, defined by (K.2);
$\boldsymbol{e}$	1	strain tensor;
e, $\boldsymbol{e}$		error and error vector, respectively;
$\boldsymbol{e}_i$	1	ith component of base vector;
$\bar{\boldsymbol{e}}_i$	1	ith component of tangent coordinate vector;
$\mathcal{F}$		extensive quantity;
F		surface or interface function;
f		function or intensive quantity;
f^B	L	bottom bounding surface of aquifer;
f_D	1	Darcy-Weisbach roughness coefficient;
f_μ	1	viscosity relation function;
f_N	1	Newton-Taylor roughness coefficient;
$\boldsymbol{f}_\sigma$	LT^{-2}	interfacial drag term of fluid momentum exchange;
f^T	L	top bounding surface of aquifer;
$\boldsymbol{f}_\tau$	$ML^{-2}T^{-2}$	deviatoric drag term of fluid momentum exchange;
G	$ML^{-1}T^{-2}$	shear modulus of the solid phase;
G	L^{-1}	first derivative of relative permeability k_r with respect to pressure head ψ;
G		number of grout zones of BHE;

G^α	L^2T^{-2}	Gibbs free energy of $\alpha-$phase;
$\boldsymbol{g}$	LT^{-2}	gravity vector;
g	LT^{-2}	$= \|\boldsymbol{g}\|$, gravitational acceleration;
$\boldsymbol{g}_i$	1	ith component of base vector of transformed coordinates;
$\bar{\boldsymbol{g}}_i$	1	ith component of tangent vector of transformed coordinates;
H	L	height;
H^α	L^2T^{-3}	supply (source/sink) of energy of $\alpha-$phase;
H_α	L^2T^{-3}	bulk source/sink term of energy of $\alpha-$phase;
H_e	$ML^{-1}T^{-3}$	overall source/sink term of internal energy;
$\bar{H}_e$	MT^{-3}	$= B\ H_e$, depth-integrated source/sink term of internal energy;
$H^\star$	L^2T^{-3}	modified heat source/sink term without well function;
h	L	hydraulic head;
h^e	L	characteristic element length or height;
I	1	ionic strength;
$\mathcal{I}$		functional;
$\boldsymbol{J}$		Jacobian matrix;
$\Im_F$	$M^{-1}LT$	Forchheimer coefficient;
$\Im_H$	$M^{-1}L^2T$	non-Fickian HC dispersion coefficient;
$\bar{\Im}_H$	$M^{-1}LT$	$= B\ \Im_H$, depth-integrated non-Fickian HC dispersion coefficient;
$\boldsymbol{j}_k$	$ML^{-2}T^{-1}$	mass flux of species k;
$\bar{\boldsymbol{j}}_k$	$ML^{-1}T^{-1}$	$= B\ \boldsymbol{j}_k$, depth-integrated mass flux of species k;
$\boldsymbol{j}_S$	$MT^{-3}\Theta^{-1}$	entropy flux;
$\boldsymbol{j}_T$	MT^{-3}	heat flux;
$\boldsymbol{j}$		arbitrary flux vector;
$\tilde{\boldsymbol{j}}$		smoothed flux;
$\boldsymbol{K}$	LT^{-1}	$= \boldsymbol{k}\rho_0 g/\mu_0$, tensor of hydraulic conductivity;
$\mathcal{K}$	ML^2T^{-2}	kinetic energy;
K_{eq}	1	equilibrium constant;
K_k^d	$M^{-1}L^3$	$= \kappa_k/\rho^s$, distribution coefficient of species k;
K_m	ML^{-3}	Michaelis-Menten's half-saturation constant;
$\boldsymbol{k}$	L^2	porous-medium permeability tensor (phase-independent);
$\boldsymbol{k}^f$	L^2	intrinsic porous-medium permeability tensor of $f-$phase;
$k_k^\dagger$	1	Langmuir numerator sorption coefficient of species k;

$k_k^{\ddagger}$	$M^{-1}L^3$	Langmuir denominator sorption coefficient of species k;
k_r	1	relative permeability;
$\boldsymbol{L}$	L^{-1}	symmetric gradient operator;
L	L	length, macroscopic scale length;
$\mathcal{L}$		differential operator, PDE (system) operator;
Le_k	1	Lewis number of species k;
ℓ		mesh level or refinement level;
M		number of phases present in the system;
M	$L^{1/3}T^{-1}$	Manning roughness coefficient;
M^f	$ML^{-1}T^{-1}\Theta^{-1}$	Soret coefficient of phase f;
$\mathcal{M}$	M	mass;
$\boldsymbol{m}$	1	specific unit vector;
m	1	VG-curve fitting parameter;
m		total number of Gauss points;
m_k	M	molecular mass of species k;
N		$= \sum_\alpha N^\alpha$, total number of chemical species;
N_I^e	1	shape function of element e at node I;
N^o		number of reactants;
N^α		number of chemical species in the $\alpha-$phase;
N_{BHE}		number of nodes of single BHE;
N_{BN}		number of nodes per element;
N_{D}		number of Dirichlet nodes;
N_{DOF}		number of degrees of freedom;
N_{E}		number of finite elements;
N_{E_p}		number of finite elements agglomerated into partition p;
N_{EQ}		number of equations;
N_{P}		number of points (nodes);
N_{PA}		number of partitions of agglomerated finite elements;
N_{PAD}		number of disjoint partitions of agglomerated finite elements;
N_{S}		number of slices;
N_k	MLT^{-3}	Dufour coefficient of species k;
N_r		number of chemical reactions;
N_Σ		number of patch elements;
Nu	1	Nusselt number;
N^*		$= \sum_\alpha (N^\alpha - 1)$, essential number of chemical species;
N_{W}		number of wells;
$\boldsymbol{n}$	1	outward-directed unit normal vector;

n_k	1	concentration exponent of species k;
n	1	pore size distribution index;
n		number of Gauss points in each direction;
$\boldsymbol{P}$	LT^{-1}	vector of accretion on a phreatic surface;
P	LT^{-1}	groundwater recharge (infiltration rate) on a phreatic surface;
Pe_t	1	thermal Darcy-Péclet number;
Pg	1	grid (mesh) Péclet number;
Pr	1	Prandtl number;
p	$ML^{-1}T^{-2}$	(thermodynamic) pressure;
p_c	$ML^{-1}T^{-2}$	capillary pressure;
p_{mech}	$ML^{-1}T^{-2}$	mechanical pressure;
Q^{α}	T^{-1}	supply (source/sink) of mass of $\alpha-$phase;
Q_{α}	T^{-1}	bulk source/sink term of mass of $\alpha-$phase;
Q	T^{-1}	$= Q_h + Q_{hw}$, bulk source/sink term of flow;
$\bar{Q}$	LT^{-1}	depth-integrated bulk source/sink term of mass;
Q_{EOB}	T^{-1}	correction sink/source term of extended Oberbeck-Boussinesq approximation;
Q_n		arbitrary integral boundary balance flux;
Q_{n_h}	L^3T^{-1}	integral boundary balance flux of liquid (positive inward-directed);
Q_h	T^{-1}	supply term of flow;
Q_{hw}	T^{-1}	specific liquid sink/source function of wells;
$\bar{Q}_h$	LT^{-1}	depth-integrated supply term of flow;
$\bar{Q}_{hw}$	LT^{-1}	depth-integrated specific liquid sink/source function of wells;
Q_k	$ML^{-3}T^{-1}$	zero-order kth-species mass sink/source function;
$\bar{Q}_k$	$ML^{-2}T^{-1}$	depth-integrated zero-order kth-species mass sink/source function;
Q_{kw}	$ML^{-3}T^{-1}$	specific kth-species mass sink/source function of wells;
$\bar{Q}_{kw}$	$ML^{-2}T^{-1}$	depth-integrated kth-species mass sink/source function of wells;
Q_r	L^3T^{-1}	total refrigerant flow discharge of BHE;
Q_T	$ML^{-1}T^{-3}$	$H_e - Q_{Tw}$, overall heat source/sink term without well function;
$\bar{Q}_T$	MT^{-3}	$\bar{H}_e - \bar{Q}_{Tw}$, depth-integrated heat source/sink term without well function;
Q_{Tw}	$ML^{-1}T^{-3}$	specific heat sink/source function of wells;
$\bar{Q}_{Tw}$	MT^{-3}	depth-integrated heat sink/source function of wells;
Q_w	L^3T^{-1}	discharge of single well w (pumping rate);
$\boldsymbol{q}^f$	LT^{-1}	$= \varepsilon_f(\boldsymbol{v}^f - \boldsymbol{v}^s)$, Darcy velocity;

$\boldsymbol{q}$	LT^{-1}	$= \boldsymbol{q}^f$, Darcy velocity of fluid;
$\bar{\boldsymbol{q}}$	L^2T^{-1}	$= B\,\boldsymbol{q}$, depth-integrated Darcy velocity;
q_n		arbitrary boundary flux (positive outward-directed);
q_{n_h}	LT^{-1}	$= \boldsymbol{q} \cdot \boldsymbol{n}$, normal boundary flux of liquid (positive outward-directed);
$\bar{q}_{n_h}$	L^2T^{-1}	$= \bar{\boldsymbol{q}} \cdot \boldsymbol{n}$, depth-integrated normal boundary flux of liquid;
$q_{n_{kC}}$	$ML^{-2}T^{-1}$	normal boundary mass flux of species k (positive outward-directed);
$\bar{q}_{n_{kC}}$	$ML^{-1}T^{-1}$	depth-integrated normal boundary mass flux of species k;
q_{n_T}	MT^{-3}	normal boundary heat flux (positive outward-directed);
$\bar{q}_{n_T}$	MLT^{-3}	depth-integrated normal boundary heat flux;
q_h	LT^{-1}	prescribed Neumann boundary liquid flux;
$\bar{q}_h$	L^2T^{-1}	prescribed depth-integrated Neumann boundary liquid flux;
q_{kC}	$ML^{-2}T^{-1}$	prescribed Neumann boundary mass flux of species k of the dispersive part;
$q_{kC}^{\dagger}$	$ML^{-2}T^{-1}$	prescribed Neumann boundary mass flux of species k of the total (convective plus dispersive) part;
$\bar{q}_{kC}$	$ML^{-1}T^{-1}$	prescribed depth-integrated Neumann boundary mass flux of species k of the dispersive part;
$\bar{q}_{kC}^{\dagger}$	$ML^{-1}T^{-1}$	prescribed depth-integrated Neumann boundary mass flux of species k of the total (convective plus dispersive) part;
q_T	MT^{-3}	prescribed Neumann boundary heat flux of the dispersive part;
$q_T^{\dagger}$	MT^{-3}	prescribed Neumann boundary heat flux of the total (convective plus dispersive) part;
$\bar{q}_T$	MLT^{-3}	prescribed depth-integrated Neumann boundary heat flux of the dispersive part;
$\bar{q}_T^{\dagger}$	MLT^{-3}	prescribed depth-integrated Neumann boundary heat flux of the total (convective plus dispersive) part;
R	$L^2T^{-2}\Theta^{-1}$	$\sim$8.314 J/°K mole, molar gas constant;
R	L	radius of pipe or wellbore;
$\bar{R}$, $\bar{R}_n$	$M^{-1}L^{-2}T^3\Theta$	thermal resistance and thermal resistance of material n, respectively;
R, R_n	$M^{-1}L^{-1}T^3\Theta$	specific thermal resistance and specific thermal resistance of material n, respectively;
R_a, R_b	$M^{-1}L^{-1}T^3\Theta$	(specific) internal borehole thermal resistance and (specific) borehole thermal resistance, respectively;

R_k	$ML^{-3}T^{-1}$	$= \sum_\alpha \varepsilon_\alpha (r_k^\alpha + R_k^\alpha)$, bulk rate of reaction of species k in all phases;
$\tilde{R}_k$	$ML^{-3}T^{-1}$	$= R_k + \sum_\alpha \varepsilon_\alpha \vartheta_k C_k^\alpha = \hat{R}_k + Q_{kw} + Q_k$, deduced bulk reaction rate;
$\hat{R}_k$	$ML^{-3}T^{-1}$	$\tilde{R}_k - Q_{kw} - Q_k$, modified bulk reaction rate;
$\bar{R}_k$	$ML^{-3}T^{-1}$	$= BR_k$, depth-integrated reaction bulk rate of species k;
$\bar{\tilde{R}}_k$	$ML^{-3}T^{-1}$	$= B\tilde{R}_k$, depth-integrated reaction bulk rate;
R_k^α	$ML^{-3}T^{-1}$	heterogeneous reaction rate of species k of $\alpha-$phase;
Ra	1	Rayleigh number;
Ra_t	1	thermal Rayleigh number;
Ra_k	1	solutal Rayleigh number of species k;
$\mathrm{Re}, \mathrm{Re}_p$	1	Reynolds number and pore Reynolds number, respectively;
$\Re^D$		$D-$dimensional Euclidean space;
$\Re_k$	1	retardation factor of species k;
$\bar{\Re}_k$	L	$= B\,\Re_k$, depth-integrated retardation factor of species k;
$\acute{\Re}_k$	1	derivative term of retardation for species k;
$\bar{\acute{\Re}}_k$	L	$= B\,\acute{\Re}_k$, depth-integrated derivative term of retardation for species k;
$\mathcal{R}$		prototypical storage (retardation) coefficient;
$\acute{\mathcal{R}}$		derivative prototypical storage (retardation) coefficient;
$\boldsymbol{r}$	L	general position vector;
r_k^α	$ML^{-3}T^{-1}$	homogeneous reaction rate of species k of $\alpha-$phase;
r	L	radial coordinate or radius;
r_b	L	borehole radius;
r_{hydr}	L	hydraulic radius;
r_r	$ML^{-3}T^{-1}$	bulk rate of reaction r;
$\mathcal{S}$	$ML^2T^{-2}\Theta^{-1}$	entropy;
S	$L^2T^{-2}\Theta^{-1}$	specific entropy density;
S	L	line boundary or segment or specific surface;
Sh_k	1	Sherwood number of species k;
S_o	L^{-1}	$= \rho_0 g(\varepsilon\gamma + \upsilon)$, specific storage coefficient;
$S_o^\star$	L^{-1}	$= \rho_0 g\varepsilon\gamma$, modified specific storage coefficient;
$\bar{S}_o$	1	$= B\,S_o$, depth-integrated specific storage coefficient;
S_k	$L^2T^{-2}\Theta^{-1}$	specific entropy density of species k;
$\boldsymbol{S}_f$	1	vector of friction slopes at channel bottom;

S_λ	1	$= 1+[(1-\varepsilon)\rho^s c^s]/(\varepsilon\rho c)$, thermal storage coefficient (thermal retardation factor);
s	1	saturation of fluid in the void space ε;
s	L	arc length of curve;
s_e	1	effective saturation;
s_p	1	pseudo-saturation;
s_r	1	residual (irreducible) saturation;
s_s	1	maximum saturation;
$\boldsymbol{T}$	L^2T^{-1}	$= \boldsymbol{k}B\rho_0 g/\mu_0$, tensor of transmissivity;
T	Θ	temperature;
T_i	Θ	inlet temperature;
T_o	Θ	outlet temperature;
T_w	Θ	prescribed temperature at well point $\boldsymbol{x}_w$;
T_*	1	tortuosity;
TBE()	L^3	total balance error;
Tu_k	1	Turner (or buoyancy) number of species k;
$\boldsymbol{t}$	1	unit tangent vector;
$\boldsymbol{t}^s$	$ML^{-1}T^{-2}$	elasticity matrix of the solid phase;
t	T	time;
t_0	T	initial time;
t_{end}	T	final time;
$t_{1/2k}$	T	reaction half-life of species k;
U_C	ML^{-3}	mass-mechanical coefficient;
U_T	$ML^{-1}T^{-2}$	thermo-mechanical coefficient;
$\boldsymbol{u}$	L	displacement;
$\boldsymbol{u}$	LT^{-1}	vector of refrigerant fluid velocity;
u	LT^{-1}	$= \Vert\boldsymbol{u}\Vert$, magnitude of refrigerant fluid velocity;
$\mathcal{V}$	MLT^{-1}	(linear) momentum;
$\boldsymbol{v}$	LT^{-1}	velocity, pore velocity;
$\boldsymbol{v}^\alpha$	LT^{-1}	velocity of the $\alpha-$phase;
$\boldsymbol{v}^{\alpha s}$	LT^{-1}	$= \boldsymbol{v}^\alpha - \boldsymbol{v}^s$, relative velocity of the $\alpha-$phase to the $s-$phase;
v	LT^{-1}	$= \Vert\boldsymbol{v}\Vert$, magnitude of velocity;
v_m	$ML^{-3}T^{-1}$	Michaelis-Menten's maximum growth rate;
W^α	$L^2T^{-3}\Theta^{-1}$	supply (source/sink) of entropy of $\alpha-$phase;
$\boldsymbol{W}$	LT^{-1}	velocity of macroscopic interface;
$W()$	1	well function;
$\boldsymbol{w}$	LT^{-1}	velocity of microscopic interface;
w	1	spatial weighting function;
w	L	width;
$\boldsymbol{X}$	L	Lagrangian material coordinates;

$\boldsymbol{x}$	L	Eulerian spatial coordinates;
$\boldsymbol{x}$	L	macroscopic position vector;
$\boldsymbol{x}_w$	L	position of well w;
x_1, x_2, x_3	L	Cartesian coordinates;
x, y, z	L	Cartesian coordinates;
$\mathcal{Z}$	L^{-1}	$= 1$ m^{-1}, unit-canceling coefficient;
z	L	axial or vertical coordinate;
$\boldsymbol{z}$	L	nodal vector of vertical coordinates;
z_k	1	charge on the kth species;

A.2 Greek Letters

α	L^{-1}	curve fitting parameter, sorptive number;
α	1	upwind parameter, $(0 \leq \alpha \leq 1)$;
α	1	specific solutal expansion coefficient of a single-species solute, density ratio;
α_k	1	specific solutal expansion coefficient of species k, density ratio;
α_L	$MT^{-2}\Theta^{-1}$	longitudinal thermodispersivity;
$\bar{\alpha}_L$	$ML^{-1}T^{-1}\Theta^{-1}$	specific longitudinal thermodispersivity;
α_T	$MT^{-2}\Theta^{-1}$	transverse thermodispersivity;
$\bar{\alpha}_T$	$ML^{-1}T^{-1}\Theta^{-1}$	specific transverse thermodispersivity;
β	Θ^{-1}	thermal expansion coefficient;
β_{c_k}	$M^{-1}L^3$	$= \alpha_k/(C_{ks} - C_{k0})$, solutal expansion coefficient of species k;
β_L	L	longitudinal dispersivity;
β_T	L	transverse dispersivity;
β_{TH}	L	horizontal transverse dispersivity;
β_{TV}	L	vertical transverse dispersivity;
Γ	L^2	areal boundary;
γ_α	1	$\alpha-$phase distribution function;
γ^f	$M^{-1}LT^2$	compressibility of $f-$phase;
γ_k	1	activity coefficient of species k;
$\boldsymbol{\Delta}$	1	Boolean matrix;
$\delta(\boldsymbol{r})$	L^{-D}	$=\begin{cases} +\infty, & \boldsymbol{r} = 0 \\ 0, & \boldsymbol{r} \neq 0 \end{cases}$, Dirac delta function of spatial dimension of $\boldsymbol{r}$, $D = 1, 2, 3$;
δ	1	error tolerance;
δ	L	microscopic scale length;
δ	1	curve fitting exponent;

δ	1	constant of friction slope relationship;
$\boldsymbol{\delta}$		unit or identity matrix;
δ_{ij}		Kronecker symbol;
$\boldsymbol{\epsilon}$	1	strain pseudovector;
$\boldsymbol{\varepsilon}$		Levi-Civita tensor;
ε	1	volume fraction, porosity (void space);
ϵ	1	error tolerance;
ϵ_2	1	(dimensionless) residual error tolerance;
ϵ_Δ	1	(dimensionless) AMR-specific error tolerance;
$\epsilon_{\|}$	1	(dimensionless) particle tracking error tolerance;
ϵ_2^*	L^3T^{-1}	(dimensional) residual error tolerance;
ε_e	1	$= \varepsilon(1 - s_r)$, specific yield (storativity of phreatic aquifer);
ε_{ijk}		permutation symbol;
$\boldsymbol{\eta}$	1	transformed (local) coordinate vector;
η	1	local coordinate;
θ		Eulerian angle;
θ	1	weighting coefficient, $(0 \le \theta \le 1)$;
θ	1	$= s\varepsilon$, moisture content;
ϑ_k	T^{-1}	decay rate of species k;
κ_k	1	Henry sorptivity coefficient of species k;
$\kappa(\boldsymbol{A})$	1	condition number of matrix $\boldsymbol{A}$;
$\boldsymbol{\Lambda}$	$MLT^{-3}\Theta^{-1}$	tensor of hydrodynamic thermodispersion;
$\bar{\boldsymbol{\Lambda}}$	$ML^2T^{-3}\Theta^{-1}$	$= B\,\boldsymbol{\Lambda}$, depth-integrated tensor of hydrodynamic thermodispersion;
$\boldsymbol{\Lambda}_0$	$MLT^{-3}\Theta^{-1}$	tensor of thermal conductivity;
$\boldsymbol{\Lambda}_{\text{mech}}$	$MLT^{-3}\Theta^{-1}$	tensor of mechanical thermodispersion;
Λ	$MLT^{-3}\Theta^{-1}$	coefficient of thermal conductivity of liquid;
Λ^α	$MLT^{-3}\Theta^{-1}$	coefficient of thermal conductivity of $\alpha-$phase;
Λ^r	$MLT^{-3}\Theta^{-1}$	coefficient of thermal conductivity of refrigerant fluid;
Λ^s	$MLT^{-3}\Theta^{-1}$	coefficient of thermal conductivity of solid;
λ^f	$ML^{-1}T^{-1}$	dilatational viscosity of the fluid phase;
λ^s	$ML^{-1}T^{-2}$	Lamé constant of the solid phase;
μ^f	$ML^{-1}T^{-1}$	dynamic viscosity of the fluid phase;
$\bar{\mu}_0^l$	$ML^{-1}T^{-1}$	specific dynamic reference viscosity of the liquid phase;
μ_k	L^2T^{-2}	chemical potential of kth-species;
μ^s	$ML^{-1}T^{-2}$	Lamé constant of the solid phase;
ν	1	Poisson's ratio;
$\nu_k,\ \nu_{kr}$	1	stoichiometric coefficient of species k (and reaction r);

ρ	ML^{-3}	mass density;
ρ_k	ML^{-3}	mass density of species k;
σ	$ML^{-3}\Theta^{-4}$	Stefan-Boltzmann constant;
$\boldsymbol{\sigma}$	$ML^{-1}T^{-2}$	stress tensor;
$\boldsymbol{\sigma}^B$	$ML^{-1}T^{-2}$	bottom surface momentum exchange vector;
$\boldsymbol{\sigma}^T$	$ML^{-1}T^{-2}$	top surface momentum exchange vector;
$\boldsymbol{\tau}$	$ML^{-1}T^{-2}$	deviatoric stress tensor;
τ		generalized friction factor;
Υ	$L^2T^{-3}\Theta^{-1}$	net production of entropy;
υ	$M^{-1}LT^2$	coefficient of skeleton compressibility;
Φ	L^2T^{-1}	potential function;
Φ_h	T^{-1}	liquid transfer (colmation/leakage) coefficient;
$\bar{\Phi}_h$	LT^{-1}	depth-integrated liquid transfer coefficient;
Φ_k	1	$= \mathrm{Le}_k/S_\lambda$, ratio of diffusivities of heat and species k;
Φ_{kC}	LT^{-1}	mass transfer coefficient of species k of the dispersive part;
$\Phi_{kC}^{\dagger}$	LT^{-1}	mass transfer coefficient of species k of the total (convective plus dispersive) part;
$\bar{\Phi}_{kC}$	L^2T^{-1}	depth-integrated mass transfer coefficient of species k of the dispersive part;
$\bar{\Phi}_{kC}^{\dagger}$	L^2T^{-1}	depth-integrated mass transfer coefficient of species k of the total (convective plus dispersive) part;
Φ_T	$MT^{-3}\Theta^{-1}$	heat transfer coefficient of the dispersive part;
$\Phi_T^{\dagger}$	$MT^{-3}\Theta^{-1}$	heat transfer coefficient of the total (convective plus dispersive) part;
$\bar{\Phi}_T$	$MLT^{-3}\Theta^{-1}$	depth-integrated heat transfer coefficient of the dispersive part;
$\bar{\Phi}_T^{\dagger}$	$MLT^{-3}\Theta^{-1}$	depth-integrated heat transfer coefficient of the total (convective plus dispersive) part;
ϕ		arbitrary function, azimuthal coordinate or Eulerian angle;
φ_k	1	adsorption function of species k;
Ξ	1	$= \Delta t_{n+1}/\Delta t_n$, maximum rate of time step change;
Ξ_{aniso}	1	ratio of anisotropy of hydraulic conductivities;
$\Xi_{\mathrm{aniso}}^{\Lambda}$	1	anisotropy factor of solid-phase thermal conductivities;
χ	1	$= (\rho - \rho_0)/\rho_0$, buoyancy coefficient;
Ψ	L^2T^{-1},L^3T^{-1}	streamfunction for 2D and axisymmetric problems, respectively;
ψ		scalar variable or Eulerian angle;
ψ	L	pressure head;

ψ_a	L	air-entry pressure head;
ψ_c	L	capillary fringe pressure head;
$\varOmega$	L^3	domain or volume;
$\boldsymbol{\omega}$	T^{-1}	vorticity;
ω	1	$\in (0,1)$, coating factor;
ω_k	1	mass fraction or specific density of species k;
ξ	1	local coordinate;
ζ	1	local coordinate;
∇	L^{-1}	Nabla (vector) operator;

A.3 Subscripts

0	reference or initial;
α	phase index taking a value of l, g, and s; subscript designates bulk quantities;
α_k	phase which contains the species k;
β	phase index taking a value of l, g, and s; subscript designates bulk quantities;
C	Cauchy-type boundary;
D	Dirichlet-type boundary;
$\varDelta$	reversal;
eq	equilibrium;
ex	external;
F	discrete feature;
f	fluid phase taking a value of l, and g; subscript designates bulk quantities;
g	gaseous phase; subscript designates bulk quantities;
g	grout;
γ	phase index taking a value of l, and g; subscript designates bulk quantities;
H	coarse mesh level;
h	fine mesh level;
het	heterogeneous;
hom	homogeneous;
int	intermediate;
i, j, l, d	spacial coordinate or global node indices;
i	pipe-in or internal;
I, J	local node indices;
k, m, n	species indicator;
k	BHE component index;

n	time plane;
n	layer, material;
L	longitudinal;
l	liquid phase; subscript designates bulk quantities;
m	Michaelis-Menten;
mech	mechanical;
N	Neumann-type boundary;
o	pipe-out or outer;
s	solid phase; subscript designates bulk quantities;
P	porous medium;
R	Robin-type boundary;
r	reaction or refrigerant;
T	transverse;
w	well;

A.4 Superscripts

ae	aerobic;
anae	anaerobic;
α	phase index taking a value of l, g, and s; superscript designates intrinsic quantities;
B	bottom;
β	phase index taking a value of l, g, and s; superscript designates intrinsic quantities;
crit	critical;
d	drying;
diff	diffusive;
e	element counter;
eff	effective;
f	fluid phase taking a value of l, and g; superscript designates intrinsic quantities;
fw	freshwater;
γ	phase index taking a value of l, and g; superscript designates intrinsic quantities;
g	gaseous phase; superscript designates intrinsic quantities;
g	grout;
g	global;
H	coarse mesh level;
h	fine mesh level;
I	interface;
i	internal;

in	inward;
l	liquid phase; superscript designates intrinsic quantities;
K	iteration number to restart;
k	iteration counter;
m	Michaelis-Menten;
o	outer;
opt	optimal;
out	outward;
π	pipe (BHE) system;
p	predictor;
r	refrigerant;
s	solid phase; superscript designates intrinsic quantities;
s	slice or soil;
sw	saltwater;
T	transpose;
T	top;
TB	top and bottom;
w	wetting;
τ	iteration counter;
$\triangle$	Delta configuration;
∇	gradient type;
Υ	Y configuration;
$+$	forward reaction;
$-$	backward reaction;

A.5 Special Symbols

$() \cdot ()$		scalar (dot) product;
$() \times ()$		vector (cross) product or Cartesian product of sets;
$() \otimes ()$		tensor (dyadic) product;
$\dot{()}$	T^{-1}	$= \partial/\partial t$, differentiation with respect to time t;
$\ddot{()}$	T^{-2}	$= \partial^2/\partial t^2$, 2nd derivative with respect to t;
$\dddot{()}$	T^{-3}	$= \partial^3/\partial t^3$, 3rd derivative with respect to t;
$()'$	L^{-1}	$= \partial/\partial x$, differentiation with respect to the 1D space coordinate x;
$()''$	L^{-2}	$= \partial^2/\partial x^2$, 2nd derivative with respect to x;
$()'''$	L^{-3}	$= \partial^3/\partial x^3$, 3rd derivative with respect to x;
$()^{(n)}$	L^{-n}	$= \partial^n/\partial x^n$, $(n = 4, 5, \ldots)$, nth derivative with respect to x;
$\hat{()}$		normalized or approximate quantity;

$(\bar{\ })^{\alpha}$		intrinsic mass average of $\alpha-$phase;
$\langle\ \rangle^{\alpha}$		intrinsic volume average of $\alpha-$phase;
$\langle\ \rangle_{\alpha}$		volume average of $\alpha-$phase;
∇	L^{-1}	gradient vector;
$\hat{\nabla}$	1	dimensionless gradient vector;
∇_i	L^{-1}	$= \partial/\partial x_i$ $(i = 1, 2, 3)$, gradient operator in x_i-coordinate direction;
∇_z	L^{-1}	$= \partial/\partial z$, gradient operator in $z-$coordinate direction;
$\nabla_{(\xi,\eta,\zeta)}$	1	gradient vector with respect to local coordinates, defined by (K.1);
$\partial/\partial t$	T^{-1}	time derivative;
$D^{\alpha}/Dt()$	T^{-1}	$= \partial()/\partial t + \boldsymbol{v}^{\alpha} \cdot \nabla()$, Eulerian material derivative of $\alpha-$phase;
$\sum$		summation notation;
$\prod$		product notation;
$\|\ \|$		absolute value of scalar or determinant of matrix;
$\|\ \|$		vector norm;
$\|\ \|_{\mathrm{RMS}}$		RMS error norm;
$\|\ \|_{L_\infty}$		maximum error norm;
$[\ ]$		molarity;
$\lceil\ \rfloor$		diagonal matrix (tensor);
$\{\ \}$		activity;

Appendix B
Coleman and Noll Method

The functional dependence for the $N(2 + D) + M(2 + 2D + D^2)$ constitutive variables listed in (3.105) is chosen as (3.106). This functional form is restricted in that the assumed dependence (3.106) of (3.105) may not violate the entropy inequality (3.69) for any process. This is the object of the Coleman and Noll method [94]. Using (3.103) the Clausius-Duhem inequality (3.69) can be written as

$$
\begin{aligned}
\rho\Upsilon = -\sum_{\alpha}\varepsilon_{\alpha}\Big\{\frac{1}{T^{\alpha}}\Big[\rho^{\alpha}\Big(\frac{D^s A^{\alpha}}{Dt} + S^{\alpha}\frac{D^s T^{\alpha}}{Dt} - \sum_{k}^{N^{\alpha}}(\mu_k^{\alpha}\frac{D^s\omega_k^{\alpha}}{Dt}) &+ \\
\boldsymbol{v}^{\alpha s}\cdot\big(\nabla A^{\alpha} + S^{\alpha}\nabla T^{\alpha} - \sum_{k}^{N^{\alpha}}(\mu_k^{\alpha}\nabla\omega_k^{\alpha})\big) &+ \\
\boldsymbol{f}_{\sigma}^{\alpha}\cdot\boldsymbol{v}^{\alpha s} + \big(A^{\alpha} - \tfrac{1}{2}v^{\alpha s^2} - \sum_{k}^{N^{\alpha}}\mu_k^{\alpha}\omega_k^{\alpha}\big)(Q^{\alpha} + Q_{\text{ex}}^{\alpha})\Big) + \frac{\boldsymbol{j}_T^{\alpha}}{T^{\alpha}}\cdot\nabla T^{\alpha} &+ \\
\boldsymbol{\sigma}^{\alpha}{:}\boldsymbol{d}^{\alpha} + \sum_{k}^{N^{\alpha}}\mu_k^{\alpha}(r_k^{\alpha} + R_k^{\alpha}) + \sum_{k}^{N^{\alpha}}\boldsymbol{j}_k^{\alpha}\cdot\nabla\mu_k^{\alpha} - \frac{\nabla T^{\alpha}}{T^{\alpha}}\cdot\sum_{k}^{N^{\alpha}}\boldsymbol{j}_k^{\alpha}\mu_k^{\alpha}\Big] &+ \\
\rho^{\alpha}W_{\text{ex}}^{\alpha}\Big\} \geq 0 &
\end{aligned}
\tag{B.1}
$$

where the interface relation (3.42) with Table 3.4 have been applied to introduce the relative velocity $\boldsymbol{v}^{\alpha s}$. The differentials in (B.1) will be developed for the chosen independent variables (3.106). For instance, the use of the chain rule to $D^s A^{\alpha}/Dt$ yields:

$$
\frac{D^s A^{\alpha}}{Dt} = \sum_{f}\frac{\partial A^{\alpha}}{\partial\varepsilon_f}\frac{D^s\varepsilon_f}{Dt} + \sum_{\beta}\frac{\partial A^{\alpha}}{\partial\rho^{\beta}}\frac{D^s\rho^{\beta}}{Dt} + \sum_{f}\frac{\partial A^{\alpha}}{\partial\boldsymbol{v}^{fs}}\cdot\frac{D^s\boldsymbol{v}^{fs}}{Dt} +
$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

$$\sum_f \frac{\partial A^\alpha}{\partial \boldsymbol{d}^f} : \frac{D^s \boldsymbol{d}^f}{Dt} + \frac{\partial A^\alpha}{\partial \boldsymbol{\epsilon}^s} \cdot \frac{D^s \boldsymbol{\epsilon}^s}{Dt} + \sum_\beta \sum_k^{N^\alpha} \frac{\partial A^\alpha}{\partial \omega_k^\beta} \frac{D^s \omega_k^\beta}{Dt} +$$

$$\sum_\beta \sum_k^{N^\alpha} \frac{\partial A^\alpha}{\partial \nabla \omega_k^\beta} \cdot \frac{D^s \nabla \omega_k^\beta}{Dt} + \sum_\beta \frac{\partial A^\alpha}{\partial T^\beta} \frac{D^s T^\beta}{Dt} + \sum_\beta \frac{\partial A^\alpha}{\partial \nabla T^\beta} \cdot \frac{D^s \nabla T^\beta}{Dt} \quad (\beta = s, f) \tag{B.2}$$

The same expansion takes place for ∇A^α:

$$\nabla A^\alpha = \sum_f \frac{\partial A^\alpha}{\partial \varepsilon_f} \nabla \varepsilon_f + \sum_\beta \frac{\partial A^\alpha}{\partial \rho^\beta} \nabla \rho^\beta + \sum_f \frac{\partial A^\alpha}{\partial \boldsymbol{v}^{fs}} \cdot \nabla(\boldsymbol{v}^{fs}) +$$

$$\sum_f \frac{\partial A^\alpha}{\partial \boldsymbol{d}^f} : \nabla(\boldsymbol{d}^f) + \frac{\partial A^\alpha}{\partial \boldsymbol{\epsilon}^s} \cdot \nabla(\boldsymbol{\epsilon}^s) + \sum_\beta \sum_k^{N^\alpha} \frac{\partial A^\alpha}{\partial \omega_k^\beta} \nabla \omega_k^\beta +$$

$$\sum_\beta \sum_k^{N^\alpha} \frac{\partial A^\alpha}{\partial \nabla \omega_k^\beta} \cdot \nabla(\nabla \omega_k^\beta) + \sum_\beta \frac{\partial A^\alpha}{\partial T^\beta} \nabla T^\beta + \sum_\beta \frac{\partial A^\alpha}{\partial \nabla T^\beta} \cdot \nabla(\nabla T^\beta) \quad (\beta = s, f) \tag{B.3}$$

and similar to $\nabla \mu_k^\alpha$. Furthermore, the conservation equation (3.48) for mass of the $\alpha-$phase ($\alpha = f, s$) can be written with (3.103) as

$$\varepsilon_\alpha \frac{D^s \rho^\alpha}{Dt} = -\boldsymbol{v}^{\alpha s} \cdot \nabla(\varepsilon_\alpha \rho^\alpha) - \varepsilon_\alpha \rho^\alpha (\boldsymbol{\delta} : \boldsymbol{d}^\alpha) + \varepsilon_\alpha \rho^\alpha (Q^\alpha + Q_{\text{ex}}^\alpha) - \rho^\alpha \frac{D^s \varepsilon_\alpha}{Dt} \tag{B.4}$$

Equation (B.4) is substituted into (B.2). Then substitution of (B.2) and (B.3) into (B.1) yields:

$$\rho \Upsilon = \sum_f \boldsymbol{v}^{fs} \cdot \Bigg\{ \sum_\alpha \Big(\frac{\varepsilon_\alpha \rho^\alpha}{T^\alpha} \frac{\partial A^\alpha}{\partial \rho^f} \frac{\rho^f}{\varepsilon_f} \nabla \varepsilon_f \Big) - \frac{\varepsilon_f \rho^f}{T^f} \Big(\sum_\gamma \Big(\frac{\partial A^\gamma}{\partial \varepsilon_\gamma} \nabla \varepsilon_\gamma \Big) +$$

$$\sum_\alpha \Big(\frac{\partial A^f}{\partial C_k^\alpha} \nabla C_k^\alpha \Big) + \sum_\alpha \Big(\frac{\partial A^f}{\partial T^\alpha} \nabla T^\alpha \Big) + S^f \nabla T^f - \sum_k^{N^f} (\mu_k^f \nabla \omega_k^f) + \boldsymbol{f}_\sigma^f \Big) \Bigg\}$$

$$- \sum_\alpha \Bigg\{ \frac{\varepsilon_\alpha \rho^\alpha}{T^\alpha} \Big(A^\alpha - \tfrac{1}{2} v^{\alpha s^2} - \sum_k^{N^\alpha} \mu_k^\alpha \omega_k^\alpha \Big) + \sum_\beta \frac{\varepsilon_\beta \rho^\beta}{T^\beta} \Big(\frac{\partial A^\beta}{\partial \rho^\alpha} \rho^\alpha \Big) \Bigg\} (Q^\alpha + Q_{\text{ex}}^\alpha)$$

$$- \sum_\beta \nabla T^\beta \cdot \Big\{ \frac{\varepsilon_\beta}{T^{\beta^2}} \big(\boldsymbol{j}_T^\beta - \sum_k^{N^\beta} \boldsymbol{j}_k^\beta \mu_k^\beta \big) + \frac{\varepsilon_\beta}{T^\beta} \sum_k^{N^\alpha} \boldsymbol{j}_k^\alpha \frac{\partial \mu_k^\alpha}{\partial T^\beta} \Big\}$$

$$-\sum_f \boldsymbol{d}^f:\Big\{\frac{\varepsilon_f\boldsymbol{\sigma}^f}{T^f} - \sum_\alpha\Big(\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\frac{\partial A^\alpha}{\partial\rho^f}\rho^f\boldsymbol{\delta}\Big) + \sum_\gamma\Big(\frac{\varepsilon_\gamma\rho^\gamma}{T^\gamma}\boldsymbol{v}^{\gamma s}\otimes\frac{\partial A^\gamma}{\partial\boldsymbol{v}^{fs}}\Big)+$$

$$\sum_\alpha\Big(\frac{\varepsilon_\alpha}{T^\alpha}\sum_k^{N^\alpha}\big(\boldsymbol{j}_k^\alpha\otimes\frac{\partial\mu_k^\alpha}{\partial\boldsymbol{v}^{fs}}\big)\Big)\Big\}$$

$$-\sum_\beta\sum_k\nabla\omega_k^\beta\cdot\Big(\sum_\alpha\frac{\varepsilon_\alpha}{T^\alpha}\boldsymbol{j}_k^\alpha\frac{\partial\mu_k^\alpha}{\partial\omega_k^\beta}\Big)$$

$$-\sum_\alpha\Big\{\varepsilon_\alpha\sum_k^{N^\alpha}\mu_k^\alpha(r_k^\alpha+R_k^\alpha)+\varepsilon_\alpha\rho^\alpha W_{\mathrm{ex}}^\alpha\Big\}$$

$$-\boldsymbol{d}^s:\Big[\frac{\varepsilon_s\boldsymbol{\sigma}^s}{T^s} - \sum_\alpha\Big(\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\frac{\partial A^\alpha}{\partial\rho^s}\rho^s\boldsymbol{\delta}\Big) + \sum_\alpha\Big(\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\frac{\partial A^\alpha}{\partial\boldsymbol{\epsilon}^s}\Big)$$

$$+\sum_f\Big(\frac{\varepsilon_f\rho^f}{T^f}\boldsymbol{v}^{fs}\otimes\sum_\gamma\frac{\partial A^f}{\partial\boldsymbol{v}^{\gamma s}}\Big) + \sum_\alpha\Big(\frac{\varepsilon_\alpha}{T^\alpha}\sum_k^{N^\alpha}\big(\boldsymbol{j}_k^\alpha\otimes\sum_\gamma\frac{\partial\mu_k^\alpha}{\partial\boldsymbol{v}^{\gamma s}}\big)\Big)\Big]$$

$$+\sum_f\frac{D^s\varepsilon_f}{Dt}\Big[-\sum_\alpha\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\Big(\frac{\partial A^\alpha}{\partial\varepsilon_f} - \frac{\rho^f}{\varepsilon_f}\frac{\partial A^\alpha}{\partial\rho^f}+\frac{\rho^s}{\varepsilon_s}\frac{\partial A^\alpha}{\partial\rho^s}\Big)\Big]$$

$$+\sum_f\nabla\varepsilon_f\cdot\Big[-\sum_\alpha\frac{\varepsilon_\alpha}{T^\alpha}\Big(\sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\frac{\partial\mu_k^\alpha}{\partial\varepsilon_f}\Big)\Big]$$

$$+\sum_f\nabla\rho^f\cdot\Big[\sum_\gamma\Big(\frac{\varepsilon_\gamma\rho^\gamma}{T^\gamma}\frac{\partial A^\gamma}{\partial\rho^f}(\boldsymbol{v}^{fs}-\boldsymbol{v}^{\gamma s})\Big)+\frac{\varepsilon_s\rho^s}{T^s}\frac{\partial A^s}{\partial\rho^f}\boldsymbol{v}^{fs}$$

$$-\sum_\alpha\frac{\varepsilon_\alpha}{T^\alpha}\Big(\sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\frac{\partial\mu_k^\alpha}{\partial\rho^f}\Big)\Big]+\sum_f\frac{D^s\boldsymbol{v}^{fs}}{Dt}\cdot\Big[-\sum_\alpha\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\frac{\partial A^\alpha}{\partial\boldsymbol{v}^{fs}}\Big]$$

$$+\sum_f\nabla(\boldsymbol{v}^{fs}):\Big[-\sum_\alpha\Big(\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\boldsymbol{v}^{\alpha s}\otimes\frac{\partial A^\alpha}{\partial\boldsymbol{v}^{fs}}\Big)-\sum_\alpha\frac{\varepsilon_\alpha}{T^\alpha}\Big(\sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\otimes\frac{\partial\mu_k^\alpha}{\partial\boldsymbol{v}^{fs}}\Big)\Big]$$

$$+\sum_f\frac{D^s\boldsymbol{d}^f}{Dt}:\Big[-\sum_\alpha\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\frac{\partial A^\alpha}{\partial\boldsymbol{d}^f}\Big]$$

$$+\sum_f\nabla(\boldsymbol{d}^f):\Big[-\sum_\alpha\Big(\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\boldsymbol{v}^{\alpha s}\otimes\frac{\partial A^\alpha}{\partial\boldsymbol{d}^f}\Big)-\sum_\alpha\frac{\varepsilon_\alpha}{T^\alpha}\Big(\sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\otimes\frac{\partial\mu_k^\alpha}{\partial\boldsymbol{d}^f}\Big)\Big]$$

$$+\sum_f\nabla(\boldsymbol{\epsilon}^s):\Big[-\sum_\alpha\Big(\frac{\varepsilon_\alpha\rho^\alpha}{T^\alpha}\boldsymbol{v}^{\alpha s}\otimes\frac{\partial A^\alpha}{\partial\boldsymbol{\epsilon}^s}\Big)-\sum_\alpha\frac{\varepsilon_\alpha}{T^\alpha}\Big(\sum_k^{N^\alpha}\boldsymbol{j}_k^\alpha\otimes\frac{\partial\mu_k^\alpha}{\partial\boldsymbol{\epsilon}^s}\Big)\Big]$$

$$
\begin{aligned}
&+\sum_{\beta}\sum_{k}^{N^{\beta}}\frac{D^s\omega_k^{\beta}}{Dt}\Big[-\sum_{\alpha}\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial\omega_k^{\beta}}+\frac{\varepsilon_{\beta}\rho^{\beta}}{T^{\beta}}\mu_k^{\beta}\Big]\\
&+\sum_{\beta}\frac{D^sT^{\beta}}{Dt}\Big[-\sum_{\alpha}\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial T^{\beta}}-\frac{\varepsilon_{\beta}\rho^{\beta}}{T^{\beta}}S^{\beta}\Big]\\
&+\sum_{\beta}\sum_{k}^{N^{\beta}}\frac{D^s\nabla\omega_k^{\beta}}{Dt}\cdot\Big[-\sum_{\alpha}\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial\nabla\omega_k^{\beta}}\Big]+\sum_{\beta}\frac{D^s\nabla T^{\beta}}{Dt}\cdot\Big[-\sum_{\alpha}\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial\nabla T^{\beta}}\Big]\\
&+\sum_{\beta}\sum_{k}^{N^{\beta}}\nabla(\nabla\omega_k^{\beta}):\Big[-\sum_{f}(\frac{\varepsilon_f\rho^f}{T^f}\boldsymbol{v}^{fs}\otimes\frac{\partial A^f}{\partial\nabla\omega_k^{\beta}})-\sum_{\alpha}\frac{\varepsilon_{\alpha}}{T^{\alpha}}(\sum_{l}^{N^l}\boldsymbol{j}_l^{\alpha}\otimes\frac{\partial\mu_l^{\alpha}}{\partial\nabla\omega_k^{\beta}})\Big]\\
&+\sum_{\beta}\nabla(\nabla T^{\beta}):\Big[-\sum_{f}(\frac{\varepsilon_f\rho^f}{T^f}\boldsymbol{v}^{fs}\otimes\frac{\partial A^f}{\partial\nabla T^{\beta}})-\sum_{\alpha}\frac{\varepsilon_{\alpha}}{T^{\alpha}}(\sum_{k}^{N^{\alpha}}\boldsymbol{j}_k^{\alpha}\otimes\frac{\partial\mu_k^{\alpha}}{\partial\nabla T^{\beta}})\Big]\\
&\geq 0\\
&\text{for}\quad(\alpha=s,f)\quad(\beta=\alpha)\quad(\gamma=f)\quad(f=1,\ldots,M-1)\\
&(k(\neq l)=1,\ldots,N^{\alpha})
\end{aligned}
\tag{B.5}
$$

The terms in the square brackets of (B.5) represent coefficients of quantities which the constitutive functions do not depend on by assumption (3.106). Therefore, the necessary and sufficient condition for $\rho\Upsilon$ to be non-negative for all independent thermodynamic states is that these coefficients have to vanish. Hence, all bracketed terms in (B.5) must be equal to zero. In doing so, the following restrictions result:

- A^{α} must be independent of $\boldsymbol{v}^{fs}$, $\boldsymbol{d}^f$, $\nabla\omega_k^{\beta}$, ∇T^{β}
- A^f is independent of $\boldsymbol{\epsilon}^s$ and $\rho^{\gamma\neq f}$
- μ_k^{α} must be independent of $\boldsymbol{v}^{fs}$, $\boldsymbol{d}^f$, $\boldsymbol{\epsilon}^s$, $\nabla\omega_k^{\beta}$, ∇T^{β}

Furthermore, we find

$$
\begin{aligned}
&\sum_{\alpha}\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial\omega_k^{\beta}}=\frac{\varepsilon_{\beta}\rho^{\beta}}{T^{\beta}}\mu_k^{\beta}\\
&\sum_{\alpha}\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial T^{\beta}}=-\frac{\varepsilon_{\beta}\rho^{\beta}}{T^{\beta}}S^{\beta}\qquad\qquad(\alpha,\beta=s,f)\\
&\sum_{\alpha}\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial\boldsymbol{\epsilon}^s}=-\frac{\varepsilon_s\boldsymbol{\sigma}^s}{T^s}+\sum_{\alpha}(\frac{\varepsilon_{\alpha}\rho^{\alpha}}{T^{\alpha}}\frac{\partial A^{\alpha}}{\partial\rho^s}\rho^s\boldsymbol{\delta})
\end{aligned}
\tag{B.6}
$$

Appendix C
Thermally Variable Fluid Density Expansion $\beta^f(T^f)$

The temperature-dependent fluid density ρ^f ranging between 0 and 100 °C can be fit by a 6th order polynomial with a high accuracy as

$$\rho^f(T^f) = a + bT^f + cT^{f^2} + dT^{f^3} + eT^{f^4} + fT^{f^5} + gT^{f^6} \quad \text{in [g/l]} \tag{C.1}$$

with the coefficients for water

$$\begin{aligned} a &= 9.998396 \cdot 10^{2} \\ b &= 6.764771 \cdot 10^{-2} \\ c &= -8.993699 \cdot 10^{-3} \\ d &= 9.143518 \cdot 10^{-5} \\ e &= -8.907391 \cdot 10^{-7} \\ f &= 5.291959 \cdot 10^{-9} \\ g &= -1.359813 \cdot 10^{-11} \end{aligned} \tag{C.2}$$

where the temperature T^f is in °C and a represents the fluid density at $T^f = 0$. Taking also into consideration of mass fraction and pressure dependencies we can argue the measured curve (C.1) is related to a reference mass fraction ω_{k0}^f and a reference pressure p_0^f. Indeed, the coefficients (C.2) have been derived for a freshwater condition $\omega_{k0}^f = 0\,(k = 1, \ldots, N^f - 1)$. Introducing a reference temperature T_0^f we can then obtain directly from (C.1) an expression for the reference fluid density ρ_0^f, viz.,

$$\begin{aligned} \rho_0^f = \rho^f(p_0^f, \omega_{k0}^f, T_0^f) = a\big|_{p_0^f, \omega_{k0}^f} + b\big|_{p_0^f, \omega_{k0}^f} T_0^f + c\big|_{p_0^f, \omega_{k0}^f} T_0^{f^2} + d\big|_{p_0^f, \omega_{k0}^f} T_0^{f^3} \\ + e\big|_{p_0^f, \omega_{k0}^f} T_0^{f^4} + f\big|_{p_0^f, \omega_{k0}^f} T_0^{f^5} + g\big|_{p_0^f, \omega_{k0}^f} T_0^{f^6} \end{aligned} \tag{C.3}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

To find the thermally variable fluid density expansion $\beta^f(T^f)$ of the fluid density function (3.199)

$$\rho^f = \rho_0^f\left[1 + \gamma^f(p^f - p_0^f) + \sum_{k=1}^{N^f-1} \alpha_k^f(\omega_k^f - \omega_{k0}^f) - \beta^f(T^f)(T^f - T_0^f)\right] \tag{C.4}$$

a Taylor series expansion for ρ^f around T_0^f, ω_{k0}^f and p_0^f is used:

$$\begin{aligned}\rho^f(T^f, \omega_k^f, p^f) = {} & \underbrace{\rho^f(T_0^f, \omega_{k0}^f, p_0^f)}_{\rho_0^f} + \frac{\partial \rho^f}{\partial T^f}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}(T^f - T_0^f) + \\ & \tfrac{1}{2}\frac{\partial^2 \rho^f}{\partial T^{f^2}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}(T^f - T_0^f)^2 + \tfrac{1}{6}\frac{\partial^3 \rho^f}{\partial T^{f^3}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}(T^f - T_0^f)^3 + \\ & \tfrac{1}{24}\frac{\partial^4 \rho^f}{\partial T^{f^4}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}(T^f - T_0^f)^4 + \tfrac{1}{120}\frac{\partial^5 \rho^f}{\partial T^{f^5}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}(T^f - T_0^f)^5 + \\ & \tfrac{1}{720}\frac{\partial^6 \rho^f}{\partial T^{f^6}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}(T^f - T_0^f)^6 + \\ & \sum_{k=1}^{N^f-1} \underbrace{\frac{\partial \rho^f}{\partial \omega_k^f}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}}_{\rho_0^f \alpha_k^f}(\omega_k^f - \omega_{k0}^f) + \underbrace{\frac{\partial \rho^f}{\partial p^f}\Big|_{T_0^f, \omega_{k0}^f, p_0^f}}_{\rho_0^f \gamma^f}(p^f - p_0^f)\end{aligned} \tag{C.5}$$

providing a 6th-order accuracy for the temperature T^f while only a linear 1st-order approximation is deemed to be sufficient for ω_k^f and p^f dependencies. By utilizing (C.1) we can evaluate the above $\partial^n \rho^f / \partial T^{f^n}$ at T_0^f, ω_{k0}^f and p_0^f according to

$$\begin{aligned}\frac{\partial \rho^f}{\partial T^f}\Big|_{T_0^f, \omega_{k0}^f, p_0^f} &= b + 2cT_0^f + 3dT_0^{f^2} + 4eT_0^{f^3} + 5fT_0^{f^4} + 6gT_0^{f^5} \\ \frac{\partial^2 \rho^f}{\partial T^{f^2}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f} &= 2c + 6dT_0^f + 12eT_0^{f^2} + 20fT_0^{f^3} + 30gT_0^{f^4} \\ \frac{\partial^3 \rho^f}{\partial T^{f^3}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f} &= 6d + 24eT_0^f + 60fT_0^{f^2} + 120gT_0^{f^3} \\ \frac{\partial^4 \rho^f}{\partial T^{f^4}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f} &= 24e + 120fT_0^f + 360gT_0^{f^2} \\ \frac{\partial^5 \rho^f}{\partial T^{f^5}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f} &= 120f + 720gT_0^f \\ \frac{\partial^6 \rho^f}{\partial T^{f^6}}\Big|_{T_0^f, \omega_{k0}^f, p_0^f} &= 720g\end{aligned} \tag{C.6}$$

As the result the fluid density ρ^f (C.5) yields

$$\begin{aligned}\rho^f(T^f, \omega_k^f, p^f) = \rho_0^f \big[1 + \sum_{k=1}^{N^f-1} \alpha_k^f(\omega_k^f - \omega_{k0}^f) + \gamma^f(p^f - p_0^f)\big]\\ +(b + 2cT_0^f + 3dT_0^{f^2} + 4eT_0^{f^3} + 5fT_0^{f^4} + 6gT_0^{f^5})(T^f - T_0^f)\\ +(c + 3dT_0^f + 6eT_0^{f^2} + 10fT_0^{f^3} + 15gT_0^{f^4})(T^f - T_0^f)^2\\ +(d + 4eT_0^f + 10fT_0^{f^2} + 20gT_0^{f^3})(T^f - T_0^f)^3\\ +(e + 5fT_0^f + 15gT_0^{f^2})(T^f - T_0^f)^4\\ +(f + 6gT_0^f)(T^f - T_0^f)^5\\ +g(T^f - T_0^f)^6 \end{aligned} \tag{C.7}$$

From (C.7) to (C.4) we finally obtain the expression for computing the nonlinear *thermally variable density expansion* to be used in (C.4)

$$\begin{aligned}\beta^f(T^f) = -\frac{1}{\rho_0^f}\Big[(b + 2cT_0^f + 3dT_0^{f^2} + 4eT_0^{f^3} + 5fT_0^{f^4} + 6gT_0^{f^5})\\ +(c + 3dT_0^f + 6eT_0^{f^2} + 10fT_0^{f^3} + 15gT_0^{f^4})(T^f - T_0^f)\\ +(d + 4eT_0^f + 10fT_0^{f^2} + 20gT_0^{f^3})(T^f - T_0^f)^2\\ +(e + 5fT_0^f + 15gT_0^{f^2})(T^f - T_0^f)^3\\ +(f + 6gT_0^f)(T^f - T_0^f)^4\\ +g(T^f - T_0^f)^5\Big] \end{aligned} \tag{C.8}$$

with ρ_0^f computed from (C.3) and the coefficients (a, b, c, d, e, f, g) as given by (C.2). As the consequence of the nonlinear thermally variable fluid density expansion the term $(1/\rho^f)\partial\rho^f/\partial T^f$ in (3.197) is no longer a constant β^f. Instead, we have

$$\frac{1}{\rho^f}\frac{\partial\rho^f}{\partial T^f} = -\frac{\beta^f(T^f) + \frac{\partial\beta^f(T^f)}{\partial T^f}(T^f - T_0^f)}{1 + \gamma^f(p^f - p_0^f) + \sum_{k=1}^{N^f-1}\alpha_k^f(\omega_k^f - \omega_{k0}^f) - \beta^f(T^f)(T^f - T_0^f)} \tag{C.9}$$

where $\beta^f(T^f)$ is taken from (C.8) and the derivation of $\beta^f(T^f)$ in (C.9) becomes

$$\begin{aligned}\frac{\partial\beta^f(T^f)}{\partial T^f} = -\frac{1}{\rho_0^f}\Big[&(c + 3dT_0^f + 6eT_0^{f^2} + 10fT_0^{f^3} + 15gT_0^{f^4})\\ &+2(d + 4eT_0^f + 10fT_0^{f^2} + 20gT_0^{f^3})(T^f - T_0^f)\\ &+3(e + 5fT_0^f + 15gT_0^{f^2})(T^f - T_0^f)^2\\ &+4(f + 6gT_0^f)(T^f - T_0^f)^3\\ &+5g(T^f - T_0^f)^4\Big]\end{aligned} \tag{C.10}$$

Appendix D
Parametric Models for Variably Saturated Porous Media

D.1 Definitions

For the sake of simplicity the fluid/liquid phase indices f, l will be omitted for the symbols used here. We introduce the *capillary pressure head* ψ of the liquid phase defined as

$$\psi = \frac{p_c}{\rho_0 g} \tag{D.1}$$

where p_c is the capillary pressure (3.221) or (3.257), ρ_0 is the reference density of the liquid phase and g is the gravitational acceleration. In an unsaturated water-air system ψ is usually negative, so ψ is sometimes termed as *suction*. Furthermore, we define the *effective saturation* of the liquid phase as

$$s_e = \frac{s - s_r}{s_s - s_r} \tag{D.2}$$

where $s_r \geq 0$ is the residual (or irreducible) saturation of liquid and s_s is the maximum saturation of liquid. We note that s_s is usually unity.

D.2 Analytic Saturation $s(\psi)$ and Inverse Capillary Pressure $\psi(s)$–Relations

D.2.1 Van Genuchten (VG) Relationship

Van Genuchten [539] proposed the analytic function

$$s_e = \begin{cases} \dfrac{1}{(1 + |\alpha\psi|^n)^m} & \text{for} \quad \psi < 0 \\ 1 & \text{for} \quad \psi \geq 0 \end{cases} \tag{D.3}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

which has gained wide acceptance in practice, where α, m and n are positive curve VG-fitting parameters. The parameter n is known as the *pore size distribution index*. Equation (D.3) is used to express the relation $s(\psi)$

$$s = s_r + (s_s - s_r)\big(1 + |\alpha\psi|^n\big)^{-m} \tag{D.4}$$

and its inverse $\psi(s)$

$$\psi = -\frac{1}{\alpha}\Big(s_e^{-\frac{1}{m}} - 1\Big)^{\frac{1}{n}} \tag{D.5}$$

Their first derivatives give[1]

$$\begin{aligned} C &= \frac{\partial s}{\partial \psi} = \frac{m\,n\,\alpha\,|\alpha\psi|^{n-1}}{(1 + |\alpha\psi|^n)^{m+1}}(s_s - s_r) \\ C^{-1} &= \frac{\partial \psi}{\partial s} = \frac{1}{m\,n\,\alpha\,(s_s - s_r)}\big(s_e^{-\frac{1}{m}} - 1\big)^{\frac{1}{n}-1}\,(s_e)^{-(1+\frac{1}{m})} \end{aligned} \tag{D.6}$$

where C and C^{-1} are known as the *moisture capacity* and the *inverse moisture capacity*, respectively.

D.2.2 Brooks-Corey Relationship

Brooks and Corey [58] and Corey [100] introduced the relationship

$$s_e = \begin{cases} \dfrac{1}{|\alpha\psi|^n} & \text{for} \quad \psi < -\frac{1}{\alpha} \\ 1 & \text{for} \quad \psi \geq -\frac{1}{\alpha} \end{cases} \tag{D.7}$$

where α and n represent positive curve fitting coefficients. The limit $(-\frac{1}{\alpha})$ in (D.7) can be identified as an air-entry pressure head $\psi_a = -\frac{1}{\alpha}$. Since

$$s = s_r + (s_s - s_r)|\alpha\psi|^{-n} \tag{D.8}$$

[1] The second derivative of $s(\psi)$ with respect to ψ reads for the VG relation

$$\frac{\partial^2 s}{\partial \psi^2} = -\left\{\frac{(m+1)mn^2\alpha^2(\alpha\psi)^{2n-2}}{[1 + (\alpha\psi)^n]^{m+2}} + \frac{m(1-n)n\alpha^2(\alpha\psi)^{n-2}}{[1 + (\alpha\psi)^n]^{m+1}}\right\}(s_s - s_r).$$

As seen s is continuously differentiable at $\psi = 0$ if $n \geq 1$. However, if $1 < n < 2$, then s is not Lipschitz continuously differentiable, and the second derivative of s is infinite at $\psi = 0$. Only for $n \geq 2$ a second derivative exists at $\psi = 0$.

and its inverse

$$\psi = -\frac{1}{\alpha} s_e^{-\frac{1}{n}} \tag{D.9}$$

their first derivatives are

$$\begin{aligned} C &= \frac{\partial s}{\partial \psi} = \frac{n\,\alpha}{|\alpha\,\psi|^{n+1}}(s_s - s_r) \\ C^{-1} &= \frac{\partial \psi}{\partial s} = \frac{1}{n\,\alpha\,(s_s - s_r)}\, s_e^{-(1+\frac{1}{n})} \end{aligned} \tag{D.10}$$

D.2.3 Haverkamp Relationship

Haverkamp [528, 529] proposed the empirical equation

$$s_e = \begin{cases} \dfrac{\alpha}{\alpha + |\mathcal{Z}\psi|^\beta} & \text{for} \quad \psi < 0 \\ 1 & \text{for} \quad \psi \geq 0 \end{cases} \tag{D.11}$$

where α and β are positive curve fitting coefficients and $\mathcal{Z} \equiv 1\ \mathrm{m}^{-1}$ is a unit-canceling coefficient. With

$$s = s_r + (s_s - s_r)\alpha(\alpha + |\mathcal{Z}\psi|^\beta)^{-1} \tag{D.12}$$

and its inverse

$$\mathcal{Z}\psi = -\big[\alpha(s_e^{-1} - 1)\big]^{\frac{1}{\beta}} \tag{D.13}$$

their first derivatives yield

$$\begin{aligned} C &= \frac{\partial s}{\partial \psi} = \frac{\alpha\,\beta\,|\mathcal{Z}\psi|^{\beta-1}}{(\alpha + |\mathcal{Z}\psi|^\beta)^2}(s_s - s_r) \\ C^{-1} &= \frac{\partial \psi}{\partial s} = \frac{\alpha}{\beta\, s_e^2\,(s_s - s_r)}\big[\alpha(s_e^{-1} - 1)\big]^{\frac{1}{\beta}-1} \end{aligned} \tag{D.14}$$

D.2.4 Exponential Relationship

Gardner [185] and Rijtema [443] proposed an exponential relation in the form

$$s_e = \begin{cases} e^{\alpha(\psi - \psi_a)} & \text{for} \quad \psi < \psi_a \\ 1 & \text{for} \quad \psi \geq \psi_a \end{cases} \tag{D.15}$$

which is applicable to analytic (exact) solutions of unsaturated flow problems, e.g., [490], where α is the only positive fitting coefficient, sometimes termed as *sorptive number*, and $\psi_a \leq 0$ is the air-entry pressure head to be prescribed. Since

$$s = s_r + (s_s - s_r)e^{\alpha(\psi - \psi_a)} \tag{D.16}$$

and its inverse

$$\psi = \frac{1}{\alpha} \ln\left(\frac{s - s_r}{s_s - s_r}\right) + \psi_a \tag{D.17}$$

their first derivatives are

$$\begin{aligned} C &= \frac{\partial s}{\partial \psi} = \alpha(s_s - s_r)e^{\alpha(\psi - \psi_a)} \\ C^{-1} &= \frac{\partial \psi}{\partial s} = \frac{1}{\alpha\,(s - s_r)} \end{aligned} \tag{D.18}$$

D.2.5 Linear Relationship

A simple linear relationship can be given in the form

$$s_e = \begin{cases} \dfrac{\psi_c - \psi}{\psi_c - \psi_a} & \text{for} \quad \psi_c < \psi < \psi_a \\ 1 & \text{for} \quad \psi \geq \psi_a \\ 0 & \text{for} \quad \psi \leq \psi_c \end{cases} \tag{D.19}$$

to approximate the capillary pressure in a capillary fringe of thickness $\psi_a - \psi_c$ at a phreatic surface, where $\psi_c < \psi_a < 0$ is the capillary fringe pressure head and $\psi_a \leq 0$ is the given air-entry pressure head. With

$$s = s_r + (s_s - s_r)\frac{\psi_c - \psi}{\psi_c - \psi_a} \tag{D.20}$$

and its inverse

$$\psi = \psi_c - (\psi_c - \psi_a)s_e \tag{D.21}$$

their first derivatives are

$$\begin{aligned} C &= \frac{\partial s}{\partial \psi} = -\frac{s_s - s_r}{\psi_c - \psi_a} \\ C^{-1} &= \frac{\partial \psi}{\partial s} = -\frac{\psi_c - \psi_a}{s_s - s_r} \end{aligned} \tag{D.22}$$

D.2.6 Time-Centered Analytic Moisture Capacity Evaluation

Mass balance accuracy in the numerical modeling of unsaturated flow is affected to a large extent by the actual treatment of the moisture capacity $C = C(\psi)$. Commonly, the evaluation of C (and accordingly C^{-1}) in time t is done at the current time $C_{n+1} = C(\psi(t_{n+1}))$, where $n + 1$ corresponds to the new time plane. Alternatively, for stability reasons it can be useful to evaluate analytically C between the previous and current time stages as [297]

$$C = (1 - \omega)C_n + \omega C_{n+1} \tag{D.23}$$

where C corresponds now a time-weighted capacity, subscripts n and $n + 1$ denote previous and current time plane, respectively, and ω is a time-weighting coefficient varying between 0 and 1. The *time-centered analytic moisture capacity* results for using $\omega = \frac{1}{2}$ in (D.23).

D.3 Analytic Relative Permeability $k_r(s)$ and $k_r(\psi)$—Relations

D.3.1 Van Genuchten-Mualem (VGM) Relationship

In a statistical approach the porous medium is conceptualized as a collection of interconnected cylindrical pores, where for each pore a laminar Poiseuille-like flow is considered having a parabolic flow profile in a tube. Based on this approach, Mualem [379] derived for the relative permeability

$$\begin{aligned} k_r &= s_e^\sigma \left(\frac{f(s_e)}{f(1)} \right)^2 \quad \text{with} \\ f(x) &= \int_0^x \frac{1}{\psi(\xi)} d\xi \end{aligned} \tag{D.24}$$

where σ is a pore connectivity parameter. Applying the VG relation (D.5)–(D.24)

$$f(x) = \int_0^x \frac{1}{\psi(\xi)} d\xi = \int_0^x \alpha \left(\frac{\xi^{\frac{1}{m}}}{1 - \xi^{\frac{1}{m}}} \right)^{\frac{1}{n}} d\xi \tag{D.25}$$

we can derive a closed solution for $k_r(s)$ after some manipulations as [536]

$$k_r(s) = s_e^{\sigma} \Big[1 - (1 - s_e^{\frac{1}{m}})^m \Big]^2 \tag{D.26}$$

if the so-called *Mualem assumption* holds

$$m = 1 - \tfrac{1}{n} \tag{D.27}$$

The exponent σ in (D.26) usually set to $\sigma = \frac{1}{2}$ accounts in Mualem's interpretation for tortuosity and connectivity so that in a physical sense $\sigma > 0$. However, σ is sometimes considered as a free fitting parameter and even negative σ can frequently be determined (e.g., [416]).

Equivalently, by using the VG relation (D.3) we can express (D.26) as the function $k_r(\psi)$ in the form

$$\begin{aligned} k_r(\psi) &= (1+\varphi)^{-\frac{5m}{2}} \big[(1+\varphi)^m - \varphi^m \big]^2 \quad \text{with the auxiliary variable} \\ \varphi &= (\alpha\psi)^n \end{aligned} \tag{D.28}$$

The first derivative of $k_r(\psi)$ (D.28) with respect to ψ yields

$$\begin{aligned} G = \frac{\partial k_r}{\partial \psi} = (n-1)\alpha(\alpha\psi)^{n-1}(1+\varphi)^{-\frac{5m}{2}} \big[(1+\varphi)^m - \varphi^m \big] \cdot \\ \left[\frac{1}{(\alpha\psi)} \Big(\frac{5\varphi}{2(1+\varphi)} - 2 \Big) - \tfrac{1}{2}(1+\varphi)^{m-1} \right] \end{aligned} \tag{D.29}$$

D.3.2 Brooks-Corey Relationship

Brooks and Corey [58] and Corey [100] suggested the following function

$$k_r(s) = s_e^{\delta} \tag{D.30}$$

where s_e is given by (D.7) and δ is a prescribed exponent. Brooks and Corey expressed the exponent δ by the pore size distribution index n in the form[2]: $\delta \approx \frac{2}{n} + 3$. Using (D.7) we can write (D.30) as the function $k_r(\psi)$ according to

$$k_r(\psi) = |\alpha\psi|^{-n\delta} \tag{D.31}$$

where its first derivative with respect to ψ is

$$G = \frac{\partial k_r}{\partial \psi} = n\,\delta\,\alpha\,|\alpha\psi|^{-(n\delta+1)} \tag{D.32}$$

D.3.3 Modified van Genuchten Relationship

In contrast to the VGM parametric model consisting of the $k_r(s)$–relation (D.26) derived from the VG $\psi(s)$–function (D.5) under the Mualem assumption (D.27), it is often advantageous to formulate the relative permeability $k_r(s)$ in the same form as (D.30)

$$k_r(s) = s_e^{\delta} \tag{D.33}$$

but now in combination with the VG-retention curve (D.3), where the fitting exponent δ can be prescribed independently. A typical value for the exponent is $\delta \approx 3$ as suggested by Irmay [286]. A simple linear relationship is obtained be setting $\delta = 1$. Inserting the VG-relation (D.3) into (D.32) we obtain the modified VG $k_r(\psi)$–relation in the form

$$k_r(\psi) = (1 + |\alpha\psi|^n)^{-m\delta} \tag{D.34}$$

[2] Using the Mualem relation (D.24) for the Brooks-Corey function (D.9) it is

$$f(x) = \int_0^x \frac{1}{\psi(\xi)} d\xi = \alpha \int_0^x \xi^{\frac{1}{n}} d\xi$$

and we can derive

$$k_r(s) = s_e^{\frac{2}{n}+\sigma+2}$$

It is obvious, with $\sigma = \frac{1}{2}$ we find

$$k_r(s) = s_e^{\delta}.$$

where

$$\delta = \frac{2}{n} + \frac{5}{2}$$

which is slightly different to the Brooks-Corey exponent $\frac{2}{n} + 3$.

with its first derivative

$$G = \frac{\partial k_r}{\partial \psi} = m\, n\, \delta\, \alpha\, \frac{|\alpha\, \psi|^{n-1}}{(1 + |\alpha\, \psi|^n)^{m\delta+1}} \tag{D.35}$$

In this alternative formulation there is no need for the Mualem assumption (D.27) and the fitting parameters n and m can be used independently.

D.3.4 Haverkamp Relationship

Haverkamp [528, 529] proposed a relationship for the relative permeability $k_r(\psi)$, which has a characteristic quite similar to the corresponding retention curve (D.11), viz.,

$$k_r(\psi) = \frac{A}{A + |\mathcal{Z}\psi|^B} \tag{D.36}$$

where A and B are positive curve fitting parameters and $\mathcal{Z} \equiv 1\ \mathrm{m}^{-1}$ is a unit-canceling coefficient. The first derivative with respect to ψ gives

$$G = \frac{\partial k_r}{\partial \psi} = \frac{A\, B\, |\mathcal{Z}\psi|^{B-1}}{(A + |\mathcal{Z}\psi|^B)^2} \tag{D.37}$$

D.3.5 Exponential Relationship

For the relative permeability Gardner [185] and Rijtema [443] proposed the relation

$$k_r(s) = s_e \tag{D.38}$$

where the exponential expression (D.15) is used for the effective saturation s_e. Inserting (D.15) into (D.38) it yields

$$k_r(\psi) = e^{\alpha(\psi - \psi_a)} \tag{D.39}$$

with its first derivative

$$G = \frac{\partial k_r}{\partial \psi} = \alpha\, e^{\alpha(\psi - \psi_a)} \tag{D.40}$$

Table D.1 Analytic parametric models for retention and relative permeability curves

Parametric model	s_e	k_r	References
(1) van Genuchten-Mualem (VGM)	$\dfrac{1}{(1+\lvert\alpha\psi\rvert^n)^m}$	$\sqrt{s_e}\Big[1-(1-s_e^{\frac{1}{m}})^m\Big]^2$	[379, 536, 539]
(2) Modified van Genuchten	$\dfrac{1}{(1+\lvert\alpha\psi\rvert^n)^m}$	s_e^δ	[100, 539]
(3) Brooks-Corey	$\dfrac{1}{\lvert\alpha\psi\rvert^n}$	s_e^δ	[58, 100]
(4) Haverkamp	$\dfrac{\alpha}{\alpha+\lvert\mathcal{Z}\psi\rvert^\beta}$	$\dfrac{A}{A+\lvert\mathcal{Z}\psi\rvert^B}$	[62, 528, 529]
(5) Exponential	$e^{\alpha(\psi-\psi_a)}$	s_e	[185, 443]
(6) Linear	$\dfrac{\psi_c-\psi}{\psi_c-\psi_a}$	s_e	

D.3.6 *Linear Relationship*

For the relative permeability a linear relationship is given by

$$k_r(s) = s_e \tag{D.41}$$

if the linear expression (D.19) is used to express the effective saturation s_e. Using (D.19) in (D.41) we obtain

$$k_r(\psi) = \frac{\psi_c - \psi}{\psi_c - \psi_a} \tag{D.42}$$

with its first derivative

$$G = \frac{\partial k_r}{\partial \psi} = -\frac{1}{\psi_c - \psi_a} \tag{D.43}$$

The analytic parametric models described above are summarized in Table D.1.

D.4 Spline Approximation of Retention and Relative Permeability Curves[3]

Analytic relations for the retention (capillary pressure) and relative permeability curves are not always suitable, because they may not describe experimental data sufficiently well in certain cases. This is exemplified for a capillary-pressure-saturation

[3] This section is contributed by V. Mirnyy (DHI-WASY).

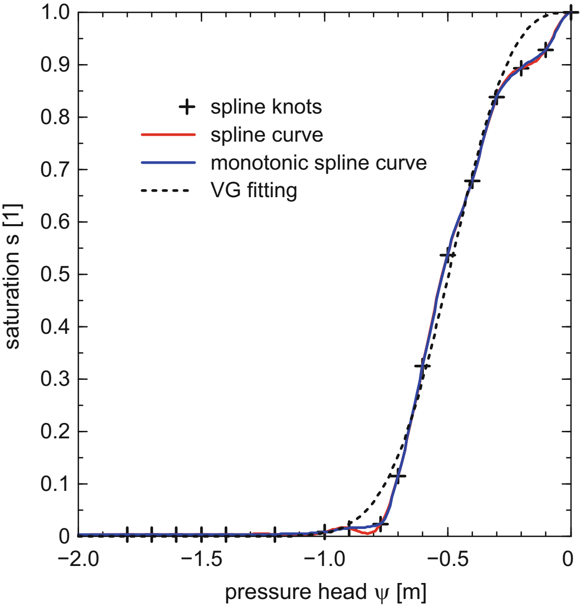


Fig. D.1 Cubic spline curves of $s(\psi)$ vs. analytic VG fitting function based on real experimental data

curve as shown Fig. D.1, where a cubic spline graph fits the experimental values very well in contrast to an analytic VG curve. The idea to automate the way from experimental data to analytic curves leads to the application of splines that can be derived directly from the experimental sample points (x_i, y_i), $i = 0, \ldots, P$ of the curves, where x_i stands for pressure head values ψ_i and saturation values s_i of the retention and relative permeability curves, respectively, and y_i stands for saturation values s_i and relative permeability values k_{ri} of the retention and relative permeability curves, respectively. Application of cubic splines to the physical characteristics of soils appeared in a couple of works during past few years. Kastanek and Nielson [303] applied classical interpolating cubic splines to the saturation-pressure relation, while allowing the introduction of 'virtual data points' to achieve necessary curve properties, such as monotonicity. Classical cubic interpolating splines that go exactly through the measured values are appropriate if the number of experimental data points is small but each measurement is of sufficiently low uncertainty (i.e., measurement errors are insignificant). In contrast to the classic cubic splines, monotonic spline approximation methods are available to enforce curve monotonicity and even positivity in the first derivatives, which has a particular relevance for the capillary pressure curve. For the following definition and analysis of different kinds of splines it is referred to [115].

Table D.2 Conditions to compute cubic interpolating spline coefficients (Note that $'$ denotes differentiation with respect to x)

Condition	Range	Number of constraints
$S_i(x_i) = y_i$	$i = 0, \dots, P-1$	P
$S_i(x_{i+1}) = y_{i+1}$	$i = 0, \dots, P-1$	P
$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$	$i = 0, \dots, P-2$	$P-1$
$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$	$i = 0, \dots, P-2$	$P-1$

D.4.1 Definition of Cubic Spline

Spline S is a piecewise polynomial continuous function defined on the interval $[x_0, x_P]$ and represents a very flexible approach of data fitting. Let the interval be covered by P disjoint subintervals $[x_i, x_{i+1}]$ with

$$a = x_0 \le x_1 \le \dots \le x_{P-1} \le x_P = b, \quad i = 0, \dots, P-1 \tag{D.44}$$

The given points $x_0, \dots, x_P$ are called *spline knots*. Interpolating spline assumes that the spline curves go exactly through the given values $y_0, \dots, y_P$ at knots $x_0, \dots, x_P$. These are the initial data for an interpolating spline:

x	x_0	x_1	$\dots$	x_P
y	y_0	y_1	$\dots$	y_P

The polynomial in the interval $[x_i, x_{i+1}]$ is defined as $S_i(x)$. Then the spline function reads

$$S(x) = \begin{cases} S_0(x) & x_0 \le x \le x_1 \\ S_1(x) & x_1 \le x \le x_2 \\ \vdots & \vdots \\ S_{P-1}(x) & x_{P-1} \le x \le x_P \end{cases} \tag{D.45}$$

Cubic polynomials are chosen in the following form

$$S_i(x) = a_i + b_i(x - x_0) + c_i(x - x_0)^2 + d_i(x - x_0)^3 \tag{D.46}$$

to achieve a continuous spline curve up to second derivative. The set of $4P$ polynomial coefficients a_i, b_i, c_i, d_i $(i = 0, \dots, P-1)$ must be determined from several conditions. Most of them are listed in Table D.2.

The first and the second row of the table describe that the polynomials S_i and S_{i+1} are connected in the point (x_{i+1}, y_{i+1}). The third and the forth row define continuity of the first and the second derivatives, respectively. Totally, the table contains $4P - 2$ conditions. Two conditions are still required to determine $4P$

polynomial coefficients uniquely. Since there are no continuity constraints in the end points, it is natural to gain the missing two conditions in the first and the last knot of the spline. There are different ways for treating BC's. Some of them are:

- Fixed slope at the end points: $S'(x_0) = C_0,\ S'(x_P) = C_P$
- Natural spline: $S''(x_0) = S''(x_P) = 0$
- Not-a-knot condition: S''' is continuous in x_1 and x_{P-1}

Table D.2 and selected BC's result in a linear algebraic system of $4P$ equations with $4P$ unknowns, which has a unique solution.

D.4.2 $s(\psi)$–Dependency

Specific BC's are employed to approximate the saturation-pressure head curve via cubic interpolating splines. We apply a natural spline condition $S''(x_0) = 0$ in the first knot and a fixed slope condition $S'(x_P) = 0$ at the last point.

D.4.2.1 Continuation to Minus Infinity

Since spline function is defined on an interval $[a, b]$, we have to advance the spline approximation to the interval $[-\infty, +\infty]$. If we claim that the last given spline knot $x_P = 0$ and $S(x_P) = 1$, then a simple constant continuation to plus infinity is

$$S_{+\infty}(x) = 1 \tag{D.47}$$

The fixed slope condition $S'(x_P) = 0$ ensures a smooth transition from S_P to $S_{+\infty}$.

In opposite direction, when $x \to -\infty$, the saturation-pressure head curve should tend to 0. It can be realized using exponential function of the form

$$S_{-\infty}(x) = ae^{bx} \tag{D.48}$$

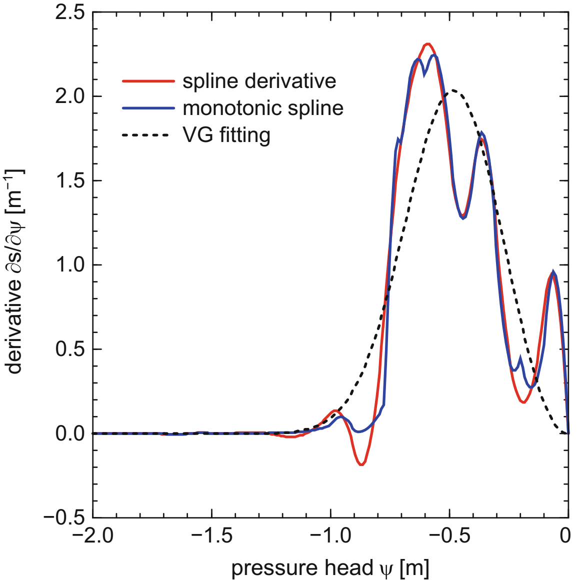
where parameters a and b can be calculated from two smoothness conditions in the first spline knot: $S_{-\infty}(x_0) = S_0(x_0)$ and $S'_{-\infty}(x_0) = S'_0(x_0)$. Defining $S_0(x_0)$ as y_0 and $S'_0(x_0)$ as m_0 we obtain

$$a = y_0 e^{-\frac{m_0 x_0}{y_0}}, \quad b = \frac{m_0}{y_0} \qquad \text{and} \qquad S_{-\infty}(x) = y_0 e^{\frac{m_0}{y_0}(x - x_0)} \tag{D.49}$$

D.4.2.2 Monotonic Spline Curve

Interpolating cubic splines described above are though continuous up to the second derivative, but may become non-monotonic even when spline knots are monotonic

Fig. D.2 Cubic spline curves of first derivatives $\partial s(\psi)/\partial\psi$ vs. analytic VG fitting function



$(y_0 \leq y_1 \leq \ldots \leq y_{P-1} \leq y_P)$. An example of such a behavior is shown in Fig. D.2, where the first derivative of the cubic spline becomes negative. For some parametric curves such fluctuations might be acceptable, while the monotonicity of the saturation-pressure head dependency is important due to its physical meaning.

Monotone interpolation can be accomplished by using cubic Hermite spline with the tangents m_i modified to ensure the monotonicity of the resulting spline. Cubic Hermit spline is generally continuous up to the first derivative only. Thus, monotonicity is achieved loosing continuity of the second derivative comparing to the cubic interpolating spline. The tangents m_i will be computed by using the Fritsch-Carlson method [176] as follows:

1. Compute the slopes of the secant lines between successive points:

$$\Delta_i = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \qquad i = 0, \ldots, P-1$$

2. Initialize the tangents at every data point as the average of the secants

$$m_i = \frac{\Delta_{i-1} + \Delta_i}{2}, \qquad i = 1, \ldots, P-1$$

For the end points one-sided differences are used: $m_0 = \Delta_0$ and $m_P = \Delta_{P-1}$.

3. If two successive points have equal values $(y_i = y_{i+1})$, then set $m_i = m_{i+1} = 0$, as the spline connecting these points must be flat to preserve monotonicity. Next steps 4 and 5 are omitted.

4. Let $\alpha_i = m_i/\Delta_i$ and $\beta_i = m_{i+1}/\Delta_i$. If $\alpha_i = 0$ or $\beta_i = 0$, then the input data points are not monotone. In such cases, piecewise monotonic curves can still be generated by choosing $m_i = m_{i+1} = 0$, although global monotonicity is not possible.
5. To prevent overshoot and ensure monotonicity, the function

$$\phi(\alpha, \beta) = \alpha - \frac{(2\alpha + \beta - 3)^2}{3(\alpha + \beta - 2)}$$

must have a value greater than zero. One simple way to satisfy this constraint is to restrict the magnitude of vector (α_i, β_i) to a circle of radius 3. That is, if $\alpha_i^2 + \beta_i^2 > 9$, then set $m_i = \tau_i \alpha_i \Delta_i$ and $m_{i+1} = \tau_i \beta_i \Delta_i$, where $\tau_i = 3/\sqrt{\alpha_i^2 + \beta_i^2}$.

For this algorithm the only one path is required.

To evaluate the cubic monotonic spline at an arbitrary point x, we find the interval $[x_k, x_{k+1}]$, such that $x_k \le x \le x_{k+1}$ by using a binary search algorithm [315]. We define, $\delta = x_{k+1} - x_k$ and $t = (x - x_k)/h$. Then the interpolant is

$$\tilde{S}(x) = y_k h_{00}(t) + \delta m_k h_{10}(t) + y_{k+1} h_{01}(t) + \delta m_{k+1} h_{11}(t) \tag{D.50}$$

where P_{ii} are the basis functions for the cubic Hermite spline:

$$\begin{aligned} h_{00}(t) &= 2t^3 - 3t^2 + 1 &&= (1 + 2t)(1 - t)^2 \\ h_{10}(t) &= t^3 - 2t^2 + t &&= t(1 - t)^2 \\ h_{01}(t) &= -2t^3 + 3t^2 &&= t^2(3 - 2t) \\ h_{11}(t) &= t^3 - t^2 &&= t^2(t - 1) \end{aligned} \tag{D.51}$$

Figure D.1 shows the classical VG fitting function for the saturation-pressure head dependency that poorly approximates experimental data given as spline knots. On the other hand, it has a well-shaped first derivative behavior (Fig. D.2). Differences between cubic interpolating and cubic monotonic spline curves are hardly visible in Fig. D.1, however, the first derivative plots (Fig. D.2) reveal clearly the positivity for the monotonic spline and the lost of the smoothness.

Appendix E
Heat Transfer and Thermal Resistance for Wall Configurations

E.1 Conduction Heat Transfer

E.1.1 Single and Composite Plane Wall

First, let us consider the thermal conduction through a single plane wall of thickness d made of solid material with a thermal conductivity Λ^s as shown in Fig. E.1a. The temperatures at the two inner and outer solid surfaces of the wall are fixed at T_1 and T_C with $T_1 > T_C$. For steady conditions without heat supply and constant thermal conductivity Λ^s, the heat transport equation, e.g., Table 3.5 (assuming $\varepsilon = 0$ for thermal conduction in solids), is for the 1D solid problem

$$-\Lambda^s \frac{\partial^2 T}{\partial x_1^2} = 0 \quad \Big(\frac{\partial}{\partial x_1} j_T = 0\Big) \tag{E.1}$$

where j_T is the Fourier heat flux in 1D, cf. (3.176). We assume that j_T is aligned with the boundary heat flux q_{n_T} normal to the wall surfaces, i.e.,

$$q_{n_T} = j_T = -\Lambda^s \frac{\partial T}{\partial x_1} \tag{E.2}$$

Using the BC's

$$T(x_1 = 0) = T_1 \quad \text{and} \quad T(x_1 = d) = T_C \tag{E.3}$$

integration of (E.1) gives the following linear temperature distribution

$$T = T_1 + (T_C - T_1)\frac{x_1}{d} \tag{E.4}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

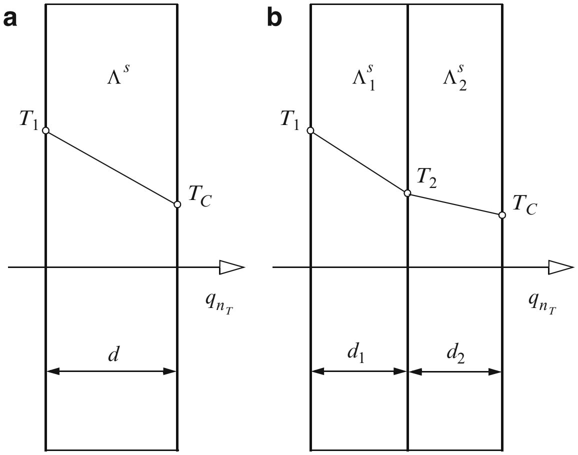


Fig. E.1 One-dimensional heat conduction through (**a**) single and (**b**) composite plane wall of solid

Applying (E.4), we obtain from (E.2) the heat flux relation

$$q_{n_T} = -\frac{\Lambda^s}{d}(T_C - T_1) \tag{E.5}$$

which corresponds to a Cauchy-type boundary flux condition, e.g., (6.40), where the heat transfer coefficient yields $\Phi_T = \Lambda^s/d$ with $T = T_1$.

Second, the same procedure is applicable to a composite wall as shown in Fig. E.1b. Due to energy conservation $\partial q_{n_T}/\partial x_1 = 0$ the heat flux q_{n_T} is constant and we find

$$q_{n_T} = -\frac{\Lambda_1^s}{d_1}(T_2 - T_1) = -\frac{\Lambda_2^s}{d_2}(T_C - T_2) \tag{E.6}$$

and eliminate T_2 to obtain

$$q_{n_T} = -\frac{1}{\left(\frac{d_1}{\Lambda_1^s} + \frac{d_2}{\Lambda_2^s}\right)}(T_C - T_1) \tag{E.7}$$

This can be further generalized to a composite wall consisting of n material layers:

$$q_{n_T} = -\frac{1}{\left(\frac{d_1}{\Lambda_1^s} + \frac{d_2}{\Lambda_2^s} + \frac{d_3}{\Lambda_3^s} + \ldots + \frac{d_n}{\Lambda_n^s}\right)}(T_C - T_1) \tag{E.8}$$

Once q_{n_T} has been determined, the temperatures at the inner material interfaces can be computed as

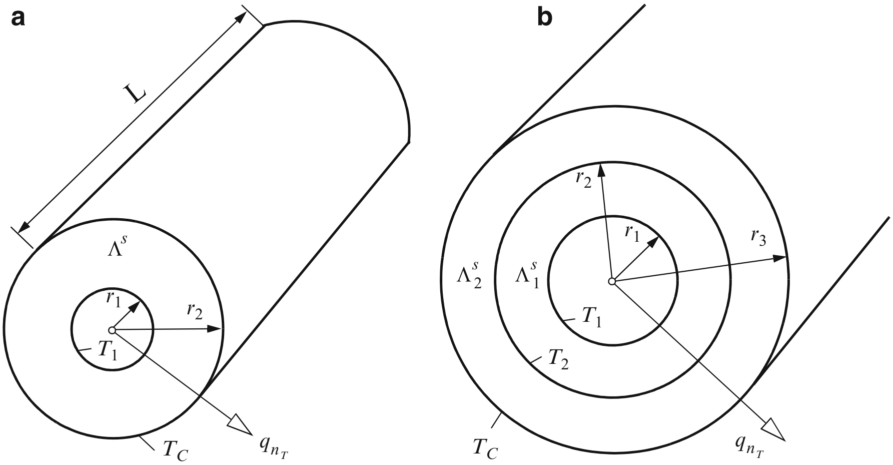


Fig. E.2 Radial heat conduction through (**a**) single and (**b**) composite circular pipe wall of solid

$$\begin{aligned} T_2 &= T_1 - q_{nT}\tfrac{d_1}{\Lambda_1^s} \\ T_3 &= T_1 - q_{nT}\big(\tfrac{d_1}{\Lambda_1^s} + \tfrac{d_2}{\Lambda_2^s}\big) \\ T_n &= T_1 - q_{nT}\big(\tfrac{d_1}{\Lambda_1^s} + \tfrac{d_2}{\Lambda_2^s} + \ldots + \tfrac{d_{n-1}}{\Lambda_{n-1}^s}\big) \end{aligned} \tag{E.9}$$

E.1.2 Single and Composite Circular Pipe Wall

First, we consider a single circular pipe wall of inside radius r_1, outside radius r_2, length L and thermal conductivity Λ^s as shown in Fig. E.2a. At the inside surface and at the outside surface of the circular pipe wall constant temperatures T_1 and T_C, respectively, are prescribed with $T_1 > T_C$. For steady-state condition without heat supply and constant thermal conductivity Λ^s, the heat transport equation reads for the axisymmetric problem (cf. Sect. 2.1.6)

$$-\Lambda^s \frac{\partial}{\partial r}\Big(r\frac{\partial T}{\partial r}\Big) = 0 \quad \Big(\frac{\partial}{\partial r} q_{nT} = 0\Big) \tag{E.10}$$

where r corresponds to the radial coordinate, supplemented by the BC's

$$T(r = r_1) = T_1 \quad \text{and} \quad T(r = r_2) = T_C \tag{E.11}$$

In the same analytical procedure as done for the single plane wall we obtain the temperature distribution for the single circular pipe wall as

$$T = T_1 + \frac{T_C - T_1}{\ln(r_2/r_1)} \ln\Big(\frac{r}{r_1}\Big) \tag{E.12}$$

which leads to the radial heat flux

$$q_{n_T} = -\frac{\Lambda^s}{r \ln(r_2/r_1)} (T_C - T_1) \tag{E.13}$$

as a function of the radial coordinate $q_{n_T} = q_{n_T}(r)$.

In analogy to the composite plane wall we find for the composite circular pipe wall as shown in Fig. E.2b the following heat flux relation:

$$q_{n_T} = -\frac{1}{r\Big(\frac{\ln(r_2/r_1)}{\Lambda_1^s} + \frac{\ln(r_3/r_2)}{\Lambda_2^s}\Big)} (T_C - T_1) \tag{E.14}$$

In generalization, the heat flux for a $n-$layered circular pipe wall system leads to

$$q_{n_T} = -\frac{1}{r\Big(\frac{\ln(r_2/r_1)}{\Lambda_1^s} + \frac{\ln(r_3/r_2)}{\Lambda_2^s} + \frac{\ln(r_4/r_3)}{\Lambda_3^s} + \ldots + \frac{\ln(r_{n+1}/r_n)}{\Lambda_n^s}\Big)} (T_C - T_1) \tag{E.15}$$

E.2 Heat Transfer Coefficient, Thermal Resistance and Specific Thermal Resistance

From above the heat flux q_{n_T} can be expressed

$$q_{n_T} = -\Phi_T (T_C - T) \tag{E.16}$$

which represents a Cauchy-type boundary heat flux condition, cf. (6.40), where T_C is the known (external or ambient) temperature on the boundary and T $(= T_1)$ is the internal temperature. It can be recognized as a specific form of the *Newton's law of cooling* [43, 447]. In (E.16) Φ_T is the *heat transfer coefficient* which is given for the configurations:

$$\Phi_T = \begin{cases} \frac{\Lambda^s}{d} & \text{single plane wall} \\ \frac{\Lambda^s}{r \ln(r_2/r_1)} & \text{single pipe wall} \\ \frac{1}{\Big(\frac{d_1}{\Lambda_1^s} + \frac{d_2}{\Lambda_2^s} + \ldots + \frac{d_n}{\Lambda_n^s}\Big)} & \text{composite plane wall} \\ \frac{1}{r\Big(\frac{\ln(r_2/r_1)}{\Lambda_1^s} + \frac{\ln(r_3/r_2)}{\Lambda_2^s} + \ldots + \frac{\ln(r_{n+1}/r_n)}{\Lambda_n^s}\Big)} & \text{composite pipe wall} \end{cases} \tag{E.17}$$

This heat transfer relation has a direct analogy to electric resistance. According to the Ohm's law the electric resistance is defined as the ratio of the voltage difference to the current flow. A *thermal resistance* $\bar{R}$ can accordingly be defined as the ratio of the temperature difference to the associated rate of heat transfer. Thus

$$\bar{R} = \frac{(T - T_C)}{\int q_{n_T} d\Gamma} = \frac{1}{\Phi_T A} \tag{E.18}$$

where $\int q_{n_T} d\Gamma$ represents the integral of the heat flux over the surface (boundary) Γ, which is approximated by

$$\int_{\Gamma} q_{n_T}\, d\Gamma \approx A\, q_{n_T} \tag{E.19}$$

where A corresponds to the exchange area. On the other hand, the *specific thermal resistance* R defines the thermal resistance per unit length and represents a material property. It is related to the heat transfer coefficient as

$$R = \frac{1}{\Phi_T S} = \bar{R}\, L \tag{E.20}$$

where $S = \frac{A}{L}$ is the specific surface and L is a length. With these definitions, the thermal resistances and specific thermal resistances of material i for a plane wall and a circular pipe wall, respectively, are

$$\bar{R}_i = \begin{cases} \dfrac{d_i}{A\, \Lambda_i^s} & \text{plane wall} \\ \dfrac{\ln(r_{i+1}/r_i)}{2\pi L\ \Lambda_i^s} & \text{circular pipe wall} \end{cases} \tag{E.21}$$

$$R_i = \begin{cases} \dfrac{d_i}{S\, \Lambda_i^s} & \text{plane wall} \\ \dfrac{\ln(r_{i+1}/r_i)}{2\pi\ \Lambda_i^s} & \text{circular pipe wall} \end{cases} \tag{E.22}$$

where for the circular pipe wall the specific exchange area is $S = 2\pi\, r$. The advantage of using specific thermal resistances is in particular for analyzing heat conduction through composite materials, where we can write for a $n-$layered structure

$$\begin{aligned} q_{n_T} &= -\frac{1}{S\, R}(T_C - T) \\ R &= R_1 + R_2 + \ldots + R_n = \textstyle\sum_{i=1}^{n} R_i \end{aligned} \tag{E.23}$$

forming a *serial* thermal resistance R composed of their partial resistance values R_i $(i = 1, \ldots, n)$ according to (E.22).

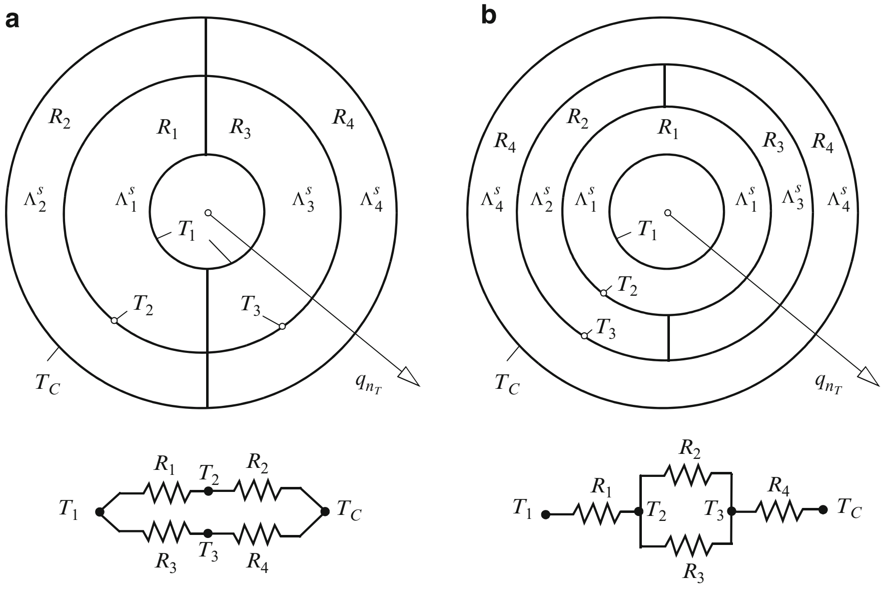


Fig. E.3 Serial and parallel heat transfer through a composite circular pipe wall with material zones showing its thermal circuit

It can also be possible that the composite wall is additionally structured in zones having different thermal properties as exemplified in Fig. E.3 for a circular pipe wall. It generates a combined serial and *parallel* heat transfer. It is usual to indicate the thermal resistance by its thermal network in direct analogy to electrical resistances in electric circuits (Fig. E.3). The serial-parallel heat transfer results for the first case (Fig. E.3a) as

$$q_{n_T} = -\frac{1}{2\,S}\left(\frac{1}{R_1 + R_2} + \frac{1}{R_3 + R_4}\right)(T_C - T) \tag{E.24}$$

and for the second case (Fig. E.3b) as

$$q_{n_T} = -\frac{1}{S\left(R_1 + \frac{1}{\frac{1}{2}\left(\frac{1}{R_2}+\frac{1}{R_3}\right)} + R_4\right)}(T_C - T) \tag{E.25}$$

where the partial thermal resistances R_i are given from (E.22). In such a way thermal resistances for more complex configurations can be developed, see e.g., [43, 237, 447].

E.3 Thermal Circuits of Prototypical Configurations in Boreholes

Borehole heat exchanger (BHE) represents a typical application of thermally interacting components of circular pipe geometries (cf. Sect. 13.5 and Appendix M) for which the effective thermal resistances can be derived by using the above technique. There are two prototypical configurations consisting of three thermal resistors: Delta configuration and Y configuration.

E.3.1 Delta Configuration

Considering a borehole cross-section containing two inner pipes denoted by subscript i (pipe-in) and subscript o (pipe-out) as shown in Fig. E.4. The borehole wall is in direct contact with the surrounding soil denoted by subscript s. In the Delta configuration (Fig. E.4a), in electrical engineering referred to as the Pi configuration, there is a series-parallel combination of the three thermal resistors between pipe-in and pipe-out, pipe-in and soil as well as pipe-out and soil, for which the equivalent specific thermal resistance can be expressed by

$$\frac{1}{R^{\vartriangle}} = \frac{1}{R_{io}} + \frac{1}{R_{is} + R_{os}} \tag{E.26}$$

or

$$R^{\vartriangle} = \frac{R_{io}(R_{is} + R_{os})}{R_{io} + R_{is} + R_{os}} \tag{E.27}$$

where R_{io}, R_{is} and R_{os} correspond to the thermal resistances of pipe-in – pipe-out, pipe-in – soil and pipe-out – soil, respectively. In the BHE context [237] the resistance R is usually termed as *internal borehole thermal resistance* $R_a = R^{\vartriangle}$ and with R_b the *borehole thermal resistance* is defined as

$$\frac{1}{R_b} = \frac{1}{R_{is}} + \frac{1}{R_{os}} \quad \text{or} \quad R_b = \frac{R_{is}R_{os}}{R_{is} + R_{os}} \tag{E.28}$$

which measures only the thermal resistance in a parallel heat transfer between the two pipes with the surrounding soil. The Delta configuration has been preferred in BHE modeling by Eskilson and Claesson [159], see Sect. 13.5 and Appendix M for more.

E.3.2 Y Configuration

The Y configuration, in electrical engineering referred to as the Wye or T configuration, describes a series combination of three thermal resistors. In the borehole

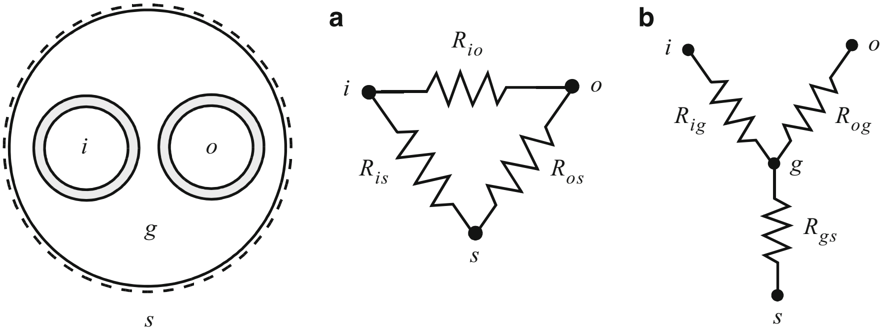


Fig. E.4 (**a**) Delta and (**b**) Y configuration for three thermal resistors in a borehole containing two inner pipes

the resistances with the backfill, the grout component denoted by subscript g, is additionally considered. As shown in Fig. E.4b for this circuit there is a serial connection of resistor R_{gs} occurring between grout and soil to a parallel connection of the resistors R_{ig} and R_{og} occurring between pip-in and grout as well as pipe-out and grout, respectively, which can be expressed by

$$\begin{aligned} R^{\curlyvee} &= R_{gs} + R_{io} \\ &= R_{gs} + \frac{1}{\frac{1}{R_{ig}} + \frac{1}{R_{og}}} \end{aligned} \tag{E.29}$$

in which R_{io} represents the equivalent thermal resistance of the parallel circuit between R_{ig} and R_{og}. The resulting thermal resistance R of (E.29) can be considered as an *internal borehole thermal resistance* $R_a = R^{\curlyvee}$ for the Y configuration. In contrast to the Delta configuration, the Y configuration entails that no direct thermal interaction between pipe-in and pipe-out exists. In BHE modeling such a Y configuration has been preferred by Al-Khoury et al. [8] and Al-Khoury and Bonnier [7].

The prototypical Delta and Y configurations represent basic circuits for BHE's, which can be enriched and extended by further inner components such as wall materials, grout zones and different pipe arrangements, so as thoroughly described in Appendix M.

Appendix F
Optimality of the Galerkin Method

The optimality of the Galerkin method indicates that the Galerkin criterion (choosing weighting function equal to the basis function) is the best approximation possible compared to any other approximation according to an appropriate measure of error. This is particularly true for 2nd-order PDE's of elliptic type and is expressed by the inequality

$$\|e\|_{E,\,\mathrm{G}} \leq \|e\|_{E,\,\mathrm{O}} \tag{F.1}$$

where $\|e\|_{E,\,\mathrm{G}}$ and $\|e\|_{E,\,\mathrm{O}}$ are the energy (Hilbert space) norm errors (8.22) produced by the Galerkin method and by any other approximation method, respectively. This error estimate is central in FEM and formulated by *Céa's lemma*, see e.g., [84, 193, 555], and tells that, at least valid for elliptic PDE's, there is never a better approximation than the Galerkin approximation. To prove (F.1) we follow Lewis et al. [345] and consider a steady-state diffusion equation of elliptic type (cf. Sect. 8.3) written in the form:

$$-\nabla \cdot (\boldsymbol{D} \cdot \nabla \phi) = H \quad \text{on} \quad \Omega \subset \Re^D \tag{F.2}$$

where ϕ is the scalar function to be solved, $\boldsymbol{D}$ is a diffusion tensor and H a source/sink term. The weak form of (F.2) is given by (cf. Sect. 8.5)

$$\int_\Omega \nabla w \cdot (\boldsymbol{D} \cdot \nabla \phi)\, d\Omega = \int_\Omega wH\, d\Omega - \int_{\Gamma_N} w q_N\, d\Gamma \tag{F.3}$$

subject to the Neumann-type BC: $q_N = -(\boldsymbol{D} \cdot \nabla\phi) \cdot \boldsymbol{n}$ on Γ_N, where $w = w(\boldsymbol{x})$ is a weighting function, which is at least once-differentiable, and $\phi = \phi(\boldsymbol{x})$ is the exact solution. The weak form (F.3) must be valid for every w.

Let us the define the error in the Galerkin approximation by

$$e_{\mathrm{G}} = \phi - \hat{\phi}_{\mathrm{G}} \tag{F.4}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

where $\hat{\phi}_{\mathrm{G}}$ is the Galerkin-based approximate solution, and the error in an approximation

$$e_{\mathrm{O}} = \phi - \hat{\phi}_{\mathrm{O}} \tag{F.5}$$

when any other approximate solution $\hat{\phi}_{\mathrm{O}}$ is used. Applying the Galerkin approximation $\phi \approx \hat{\phi}_{\mathrm{G}}$ to the weak form (F.3), it must satisfy

$$\int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla \hat{\phi}_{\mathrm{G}})\, d\Omega = \int_{\Omega} wH\, d\Omega - \int_{\Gamma_N} wq_N\, d\Gamma \tag{F.6}$$

for every w. Subtracting (F.6) from (F.3) it yields

$$\int_{\Omega} \nabla w \cdot (\boldsymbol{D} \cdot \nabla e_{\mathrm{G}})\, d\Omega = 0 \tag{F.7}$$

for every w.

Let us define the energy (Hilbert space) norm error of a function f according to (8.22) in the following form

$$\| f \|_E^2 = \int_{\Omega} \nabla f \cdot (\boldsymbol{D} \cdot \nabla f)\, d\Omega \tag{F.8}$$

we can expand (F.5) to

$$e_{\mathrm{O}} = \phi - \hat{\phi}_{\mathrm{O}} = (\phi - \hat{\phi}_{\mathrm{G}}) + (\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}) = e_{\mathrm{G}} + (\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}) \tag{F.9}$$

and find

$$\|e\|_{E,\,\mathrm{O}}^2 = \|e\|_{E,\,\mathrm{G}}^2 + 2\|(\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}})e_{\mathrm{G}}\|_E + \|\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}\|_E^2 \tag{F.10}$$

Using (F.8) and (F.10) leads to

$$\begin{aligned}\|e\|_{E,\,\mathrm{O}}^2 = \int_{\Omega} \nabla e_{\mathrm{G}} \cdot (\boldsymbol{D} \cdot \nabla e_{\mathrm{G}})\, d\Omega + 2\int_{\Omega} \nabla(\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}) \cdot (\boldsymbol{D} \cdot \nabla e_{\mathrm{G}})\, d\Omega +\\ \int_{\Omega} \nabla(\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}) \cdot \big[\boldsymbol{D} \cdot \nabla(\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}})\big]\, d\Omega\end{aligned} \tag{F.11}$$

Since (F.7) is valid for every weighting function w, the second term on the RHS of (F.11) must vanish

$$\int_{\Omega} \nabla(\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}) \cdot (\boldsymbol{D} \cdot \nabla e_{\mathrm{G}})\, d\Omega = 0 \tag{F.12}$$

and it follows from (F.11)

$$\|e\|^2_{E,\,\mathrm{O}} = \|e\|^2_{E,\,\mathrm{G}} + \|\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}\|^2_E \tag{F.13}$$

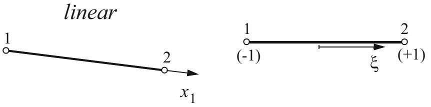
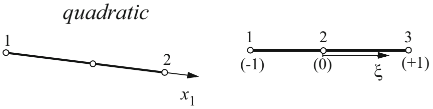
From the definition of the error measure (F.8) it results that $\|\hat{\phi}_{\mathrm{G}} - \hat{\phi}_{\mathrm{O}}\|^2_E \geq 0$ is strictly non-negative and the optimality condition (F.1) becomes now apparent.

Appendix G
Isoparametric Finite Element Shape Functions and Their Derivatives

Standard isoparametric elements in 1D, 2D and 3D can be found in finite element textbooks, e.g., [590]. Pyramidal 3D elements are described by Zgainski et al. [586].

G.1 One Dimension

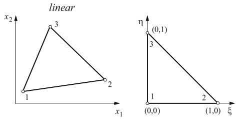
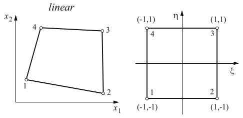
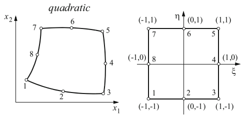
Table G.1 One-dimensional isoparametric shape functions N_I^e and their derivatives at node I of element $\bar{\Omega}^e \subset \Re^1$ of type: (a) line and (b) parabola

Type	I	N_I^e	$\frac{\partial N_I^e}{\partial \xi}$	Element in global and local coordinates
(a)	1	$\frac{1}{2}(1-\xi)$	$-\frac{1}{2}$	 *linear* 1 2 x_1 1 2 (-1) ξ (+1)
	2	$\frac{1}{2}(1+\xi)$	$\frac{1}{2}$	
(b)	1	$\frac{1}{2}\xi(\xi-1)$	$\xi - \frac{1}{2}$	 *quadratic* 1 2 x_1 1 2 3 (-1) (0) ξ (+1)
	2	$1-\xi^2$	-2ξ	
	3	$\frac{1}{2}\xi(\xi+1)$	$\xi + \frac{1}{2}$	

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,
© Springer-Verlag Berlin Heidelberg 2014

G.2 Two Dimensions

Table G.2 Two-dimensional isoparametric shape functions N_I^e and their derivatives at node I of element $\bar{\Omega}^e \subset \Re^2$ of type: (a) triangle, (b) quadrilateral and (c) curved quadrilateral

Type	I	N_I^e	$\frac{\partial N_I^e}{\partial \xi}$	$\frac{\partial N_I^e}{\partial \eta}$	Element in global and local coordinates
(a)	1	$1-\xi-\eta$	-1	-1	 linear x_2 x_1 1 2 3 η ξ (0,1) (0,0) (1,0)
	2	ξ	1	0	
	3	η	0	1	
(b)	1	$\frac{1}{4}(1-\xi)(1-\eta)$	$-\frac{1}{4}(1-\eta)$	$-\frac{1}{4}(1-\xi)$	 linear x_2 x_1 1 2 3 4 η ξ (-1,1) (1,1) (-1,-1) (1,-1)
	2	$\frac{1}{4}(1+\xi)(1-\eta)$	$\frac{1}{4}(1-\eta)$	$-\frac{1}{4}(1+\xi)$	
	3	$\frac{1}{4}(1+\xi)(1+\eta)$	$\frac{1}{4}(1+\eta)$	$\frac{1}{4}(1+\xi)$	
	4	$\frac{1}{4}(1-\xi)(1+\eta)$	$-\frac{1}{4}(1+\eta)$	$\frac{1}{4}(1-\xi)$	
(c)	1	$\frac{1}{4}(1-\xi)(1-\eta)(-\xi-\eta-1)$	$\frac{1}{4}(\eta+2\xi-2\xi\eta-\eta^2)$	$\frac{1}{4}(\xi+2\eta-\xi^2-2\xi\eta)$	 quadratic x_2 x_1 1 2 3 4 5 6 7 8 η ξ (-1,1) (0,1) (1,1) (-1,0) (1,0) (-1,-1) (0,-1) (1,-1)
	2	$\frac{1}{2}(1-\xi^2)(1-\eta)$	$-\xi(1-\eta)$	$-\frac{1}{2}(1-\xi^2)$	
	3	$\frac{1}{4}(1+\xi)(1-\eta)(\xi-\eta-1)$	$\frac{1}{4}(-\eta+2\xi-2\xi\eta+\eta^2)$	$\frac{1}{4}(-\xi+2\eta-\xi^2+2\xi\eta)$	
	4	$\frac{1}{2}(1+\xi)(1-\eta^2)$	$\frac{1}{2}(1-\eta^2)$	$-\eta(1+\xi)$	
	5	$\frac{1}{4}(1+\xi)(1+\eta)(\xi+\eta-1)$	$\frac{1}{4}(\eta+2\xi+2\xi\eta+\eta^2)$	$\frac{1}{4}(\xi+2\eta+\xi^2+2\xi\eta)$	
	6	$\frac{1}{2}(1-\xi^2)(1+\eta)$	$-\xi(1+\eta)$	$\frac{1}{2}(1-\xi^2)$	
	7	$\frac{1}{4}(1-\xi)(1+\eta)(-\xi+\eta-1)$	$\frac{1}{4}(-\eta+2\xi+2\xi\eta-\eta^2)$	$\frac{1}{4}(-\xi+2\eta+\xi^2-2\xi\eta)$	
	8	$\frac{1}{2}(1-\xi)(1-\eta^2)$	$-\frac{1}{2}(1-\eta^2)$	$-\eta(1-\xi)$	

G.3 Three Dimensions

Table G.3 Three-dimensional linear isoparametric shape functions N_I^e and their derivatives at node I of element $\bar{\Omega}^e \subset \Re^3$ of type: (a) tetrahedron, (b) triangular prism (pentahedron), (c) quadrilateral prism (hexahedron) and (d) square pyramid

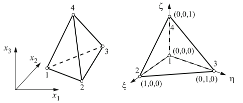
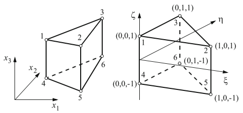
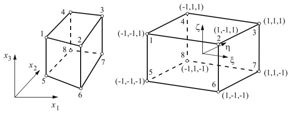
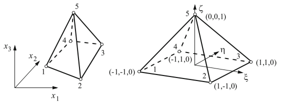
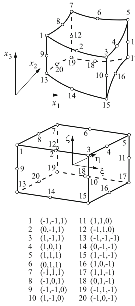
Type	I	N_I^e	$\frac{\partial N_I^e}{\partial \xi}$	$\frac{\partial N_I^e}{\partial \eta}$	$\frac{\partial N_I^e}{\partial \zeta}$	Element in global and local coordinates
(a)	1	$1-\xi-\eta-\zeta$	-1	-1	-1	 x_3 x_2 x_1 1 2 3 4 ζ (0,0,1) 4 (0,0,0) 1 2 3 ξ (1,0,0) (0,1,0) η
	2	ξ	1	0	0	
	3	η	0	1	0	
	4	ζ	0	0	1	
(b)	1	$\frac{1}{2}(1-\xi-\eta)(1+\zeta)$	$-\frac{1}{2}(1+\zeta)$	$-\frac{1}{2}(1+\zeta)$	$\frac{1}{2}(1-\xi-\eta)$	 x_3 x_2 x_1 1 2 3 4 5 6 ζ (0,1,1) 3 η (0,0,1) 1 2 (1,0,1) 6 (0,1,-1) 4 (0,0,-1) 5 ξ (1,0,-1)
	2	$\frac{1}{2}\xi(1+\zeta)$	$\frac{1}{2}(1+\zeta)$	0	$\frac{1}{2}\xi$	
	3	$\frac{1}{2}\eta(1+\zeta)$	0	$\frac{1}{2}(1+\zeta)$	$\frac{1}{2}\eta$	
	4	$\frac{1}{2}(1-\xi-\eta)(1-\zeta)$	$-\frac{1}{2}(1-\zeta)$	$-\frac{1}{2}(1-\zeta)$	$-\frac{1}{2}(1-\xi-\eta)$	
	5	$\frac{1}{2}\xi(1-\zeta)$	$\frac{1}{2}(1-\zeta)$	0	$-\frac{1}{2}\xi$	
	6	$\frac{1}{2}\eta(1-\zeta)$	0	$\frac{1}{2}(1-\zeta)$	$-\frac{1}{2}\eta$	
(c)	1	$\frac{1}{8}(1-\xi)(1-\eta)(1+\zeta)$	$-\frac{1}{8}(1-\eta)(1+\zeta)$	$-\frac{1}{8}(1-\xi)(1+\zeta)$	$\frac{1}{8}(1-\xi)(1-\eta)$	 x_3 x_2 x_1 1 2 3 4 5 6 7 8 (-1,1,1) 4 ζ (1,-1,1) (1,1,1) (-1,-1,1) 1 2 3 η 8 ξ (-1,1,-1) 7 5 (1,1,-1) (-1,-1,-1) 6 (1,-1,-1)
	2	$\frac{1}{8}(1+\xi)(1-\eta)(1+\zeta)$	$\frac{1}{8}(1-\eta)(1+\zeta)$	$-\frac{1}{8}(1+\xi)(1+\zeta)$	$\frac{1}{8}(1+\xi)(1-\eta)$	
	3	$\frac{1}{8}(1+\xi)(1+\eta)(1+\zeta)$	$\frac{1}{8}(1+\eta)(1+\zeta)$	$\frac{1}{8}(1+\xi)(1+\zeta)$	$\frac{1}{8}(1+\xi)(1+\eta)$	
	4	$\frac{1}{8}(1-\xi)(1+\eta)(1+\zeta)$	$-\frac{1}{8}(1+\eta)(1+\zeta)$	$\frac{1}{8}(1-\xi)(1+\zeta)$	$\frac{1}{8}(1-\xi)(1+\eta)$	
	5	$\frac{1}{8}(1-\xi)(1-\eta)(1-\zeta)$	$-\frac{1}{8}(1-\eta)(1-\zeta)$	$-\frac{1}{8}(1-\xi)(1-\zeta)$	$-\frac{1}{8}(1-\xi)(1-\eta)$	
	6	$\frac{1}{8}(1+\xi)(1-\eta)(1-\zeta)$	$\frac{1}{8}(1-\eta)(1-\zeta)$	$-\frac{1}{8}(1+\xi)(1-\zeta)$	$-\frac{1}{8}(1+\xi)(1-\eta)$	
	7	$\frac{1}{8}(1+\xi)(1+\eta)(1-\zeta)$	$\frac{1}{8}(1+\eta)(1-\zeta)$	$\frac{1}{8}(1+\xi)(1-\zeta)$	$-\frac{1}{8}(1+\xi)(1+\eta)$	
	8	$\frac{1}{8}(1-\xi)(1+\eta)(1-\zeta)$	$-\frac{1}{8}(1+\eta)(1-\zeta)$	$\frac{1}{8}(1-\xi)(1-\zeta)$	$-\frac{1}{8}(1-\xi)(1+\eta)$	
(d)	1	$\frac{1}{4}[(1-\xi)(1-\eta)-\zeta+\frac{\xi\eta\zeta}{1-\zeta}]$	$-\frac{1}{4}[(1-\eta)-\frac{\eta\zeta}{1-\zeta}]$	$-\frac{1}{4}[(1-\xi)-\frac{\xi\zeta}{1-\zeta}]$	$-\frac{1}{4}[1-\frac{\xi\eta}{(1-\zeta)^2}]$	 x_3 x_2 x_1 1 2 3 4 5 5 ζ (0,0,1) 4 η (-1,1,0) 3 (1,1,0) (-1,-1,0) 1 2 ξ (1,-1,0)
	2	$\frac{1}{4}[(1+\xi)(1-\eta)-\zeta-\frac{\xi\eta\zeta}{1-\zeta}]$	$\frac{1}{4}[(1-\eta)-\frac{\eta\zeta}{1-\zeta}]$	$-\frac{1}{4}[(1+\xi)+\frac{\xi\zeta}{1-\zeta}]$	$-\frac{1}{4}[1+\frac{\xi\eta}{(1-\zeta)^2}]$	
	3	$\frac{1}{4}[(1+\xi)(1+\eta)-\zeta+\frac{\xi\eta\zeta}{1-\zeta}]$	$\frac{1}{4}[(1+\eta)+\frac{\eta\zeta}{1-\zeta}]$	$\frac{1}{4}[(1+\xi)+\frac{\xi\zeta}{1-\zeta}]$	$-\frac{1}{4}[1-\frac{\xi\eta}{(1-\zeta)^2}]$	
	4	$\frac{1}{4}[(1-\xi)(1+\eta)-\zeta-\frac{\xi\eta\zeta}{1-\zeta}]$	$-\frac{1}{4}[(1+\eta)+\frac{\eta\zeta}{1-\zeta}]$	$\frac{1}{4}[(1-\xi)-\frac{\xi\zeta}{1-\zeta}]$	$-\frac{1}{4}[1+\frac{\xi\eta}{(1-\zeta)^2}]$	
	5	ζ	0	0	1	

Table G.4 Three-dimensional quadratic isoparametric shape functions N_I^e and their derivatives at node I of element $\bar{\Omega}^e \subset \Re^3$ of curved hexahedral element type

I	N_I^e	$\frac{\partial N_I^e}{\partial \xi}$	$\frac{\partial N_I^e}{\partial \eta}$	$\frac{\partial N_I^e}{\partial \zeta}$	Element in global and local coordinates
1	$\frac{1}{8}[(1-\xi)(1-\eta)(1+\zeta)$	$\frac{1}{8}[(-1+2\xi)(1-\eta)(1+\zeta)-$	$\frac{1}{8}[(-1+2\eta)(1-\xi)(1+\zeta)-$	$\frac{1}{8}[(1+2\zeta)(1-\xi)(1-\eta)+$	
	$(-\xi-\eta+\zeta-2)]$	$(1-\eta)(1+\zeta)(-\eta+\zeta-2)]$	$(1-\xi)(1+\zeta)(-\xi+\zeta-2)]$	$(1-\xi)(1-\eta)(-\xi-\eta-2)]$	
2	$\frac{1}{4}(1-\xi^2)(1-\eta)(1+\zeta)$	$-\frac{1}{2}\xi(1-\eta)(1+\zeta)$	$-\frac{1}{4}(1-\xi^2)(1+\zeta)$	$\frac{1}{4}(1-\xi^2)(1-\eta)$	
3	$\frac{1}{8}[(1+\xi)(1-\eta)(1+\zeta)$	$\frac{1}{8}[(1+2\xi)(1-\eta)(1+\zeta)+$	$\frac{1}{8}[(-1+2\eta)(1+\xi)(1+\zeta)-$	$\frac{1}{8}[(1+2\zeta)(1+\xi)(1-\eta)+$	
	$(\xi-\eta+\zeta-2)]$	$(1-\eta)(1+\zeta)(-\eta+\zeta-2)]$	$(1+\xi)(1+\zeta)(\xi+\zeta-2)]$	$(1+\xi)(1-\eta)(\xi-\eta-2)]$	
4	$\frac{1}{4}(1+\xi)(1-\eta^2)(1+\zeta)$	$\frac{1}{4}[(1-\eta^2)(1+\zeta)]$	$-\frac{1}{2}\eta(1+\xi)(1+\zeta)$	$\frac{1}{4}[(1+\xi)(1-\eta^2)]$	
5	$\frac{1}{8}[(1+\xi)(1+\eta)(1+\zeta)$	$\frac{1}{8}[(1+2\xi)(1+\eta)(1+\zeta)+$	$\frac{1}{8}[(1+2\eta)(1+\xi)(1+\zeta)+$	$\frac{1}{8}[(1+2\zeta)(1+\xi)(1+\eta)+$	
	$(\xi+\eta+\zeta-2)]$	$(1+\eta)(1+\zeta)(\eta+\zeta-2)]$	$(1+\xi)(1+\zeta)(\xi+\zeta-2)]$	$(1+\xi)(1+\eta)(\xi+\eta-2)]$	
6	$\frac{1}{4}(1-\xi^2)(1+\eta)(1+\zeta)$	$-\frac{1}{2}\xi(1+\eta)(1+\zeta)$	$\frac{1}{4}(1-\xi^2)(1+\zeta)$	$\frac{1}{4}(1-\xi^2)(1+\eta)$	
7	$\frac{1}{8}[(1-\xi)(1+\eta)(1+\zeta)$	$\frac{1}{8}[(-1+2\xi)(1+\eta)(1+\zeta)-$	$\frac{1}{8}[(1+2\eta)(1-\xi)(1+\zeta)+$	$\frac{1}{8}[(1+2\zeta)(1-\xi)(1+\eta)+$	 7 6 5 8 1 12 2 4 11 3 9 19 18 17 20 13 10 16 14 15 x_1 x_2 x_3 ζ η ξ 1 (-1,-1,1) 2 (0,-1,1) 3 (1,-1,1) 4 (1,0,1) 5 (1,1,1) 6 (0,1,1) 7 (-1,1,1) 8 (-1,0,1) 9 (-1,-1,0) 10 (1,-1,0) 11 (1,1,0) 12 (-1,1,0) 13 (-1,-1,-1) 14 (0,-1,-1) 15 (1,-1,-1) 16 (1,0,-1) 17 (1,1,-1) 18 (0,1,-1) 19 (-1,1,-1) 20 (-1,0,-1)
	$(-\xi+\eta+\zeta-2)]$	$(1+\eta)(1+\zeta)(\eta+\zeta-2)]$	$(1-\xi)(1+\zeta)(-\xi+\zeta-2)]$	$(1-\xi)(1+\eta)(-\xi+\eta-2)]$	
8	$\frac{1}{4}(1-\xi)(1-\eta^2)(1+\zeta)$	$-\frac{1}{4}(1-\eta^2)(1+\zeta)$	$-\frac{1}{2}\eta(1-\xi)(1+\zeta)$	$\frac{1}{4}(1-\xi)(1-\eta^2)$	
9	$\frac{1}{4}(1-\xi)(1-\eta)(1-\zeta^2)$	$-\frac{1}{4}(1-\eta)(1-\zeta^2)$	$-\frac{1}{4}(1-\xi)(1-\zeta^2)$	$-\frac{1}{2}\zeta(1-\xi)(1-\eta)$	
10	$\frac{1}{4}(1+\xi)(1-\eta)(1-\zeta^2)$	$\frac{1}{4}(1-\eta)(1-\zeta^2)$	$-\frac{1}{4}(1+\xi)(1-\zeta^2)$	$-\frac{1}{2}\zeta(1+\xi)(1-\eta)$	
11	$\frac{1}{4}(1+\xi)(1+\eta)(1-\zeta^2)$	$\frac{1}{4}(1+\eta)(1-\zeta^2)$	$\frac{1}{4}(1+\xi)(1-\zeta^2)$	$-\frac{1}{2}\zeta(1+\xi)(1+\eta)$	
12	$\frac{1}{4}(1-\xi)(1+\eta)(1-\zeta^2)$	$-\frac{1}{4}(1+\eta)(1-\zeta^2)$	$\frac{1}{4}(1-\xi)(1-\zeta^2)$	$-\frac{1}{2}\zeta(1-\xi)(1+\eta)$	
13	$\frac{1}{8}[(1-\xi)(1-\eta)(1-\zeta)$	$\frac{1}{8}[(-1+2\xi)(1-\eta)(1-\zeta)-$	$\frac{1}{8}[(-1+2\eta)(1-\xi)(1-\zeta)-$	$\frac{1}{8}[(-1+2\zeta)(1-\xi)(1-\eta)-$	
	$(-\xi-\eta-\zeta-2)]$	$(1-\eta)(1-\zeta)(-\eta-\zeta-2)]$	$(1-\xi)(1-\zeta)(-\xi-\zeta-2)]$	$(1-\xi)(1-\eta)(-\xi-\eta-2)]$	
14	$\frac{1}{4}(1-\xi^2)(1-\eta)(1-\zeta)$	$-\frac{1}{2}\xi(1-\eta)(1-\zeta)$	$-\frac{1}{4}(1-\xi^2)(1-\zeta)$	$-\frac{1}{4}(1-\xi^2)(1-\eta)$	
15	$\frac{1}{8}[(1+\xi)(1-\eta)(1-\zeta)$	$\frac{1}{8}[(1+2\xi)(1-\eta)(1-\zeta)+$	$\frac{1}{8}[(-1+2\eta)(1+\xi)(1-\zeta)-$	$\frac{1}{8}[(-1+2\zeta)(1+\xi)(1-\eta)-$	
	$(\xi-\eta-\zeta-2)]$	$(1-\eta)(1-\zeta)(-\eta-\zeta-2)]$	$(1+\xi)(1-\zeta)(\xi-\zeta-2)]$	$(1+\xi)(1-\eta)(\xi-\eta-2)]$	
16	$\frac{1}{4}(1+\xi)(1-\eta^2)(1-\zeta)$	$\frac{1}{4}(1-\eta^2)(1-\zeta)$	$-\frac{1}{2}\eta(1+\xi)(1-\zeta)$	$-\frac{1}{4}(1+\xi)(1-\eta^2)$	
17	$\frac{1}{8}[(1+\xi)(1+\eta)(1-\zeta)$	$\frac{1}{8}[(1+2\xi)(1+\eta)(1-\zeta)+$	$\frac{1}{8}[(1+2\eta)(1+\xi)(1-\zeta)+$	$\frac{1}{8}[(-1+2\zeta)(1+\xi)(1+\eta)-$	
	$(\xi+\eta-\zeta-2)]$	$(1+\eta)(1-\zeta)(\eta-\zeta-2)]$	$(1+\xi)(1-\zeta)(\xi-\zeta-2)]$	$(1+\xi)(1+\eta)(\xi+\eta-2)]$	
18	$\frac{1}{4}(1-\xi^2)(1+\eta)(1-\zeta)$	$-\frac{1}{2}\xi(1+\eta)(1-\zeta)$	$\frac{1}{4}(1-\xi^2)(1-\zeta)$	$-\frac{1}{4}(1-\xi^2)(1+\eta)$	
19	$\frac{1}{8}[(1-\xi)(1+\eta)(1-\zeta)$	$\frac{1}{8}[(-1+2\xi)(1+\eta)(1-\zeta)-$	$\frac{1}{8}[(1+2\eta)(1-\xi)(1-\zeta)+$	$\frac{1}{8}[(-1+2\zeta)(1-\xi)(1+\eta)-$	
	$(-\xi+\eta-\zeta-2)]$	$(1+\eta)(1-\zeta)(\eta-\zeta-2)]$	$(1-\xi)(1-\zeta)(-\xi-\zeta-2)]$	$(1-\xi)(1+\eta)(-\xi+\eta-2)]$	
20	$\frac{1}{4}(1-\xi)(1-\eta^2)(1-\zeta)$	$-\frac{1}{4}(1-\eta^2)(1-\zeta)$	$-\frac{1}{2}\eta(1-\xi)(1-\zeta)$	$-\frac{1}{4}(1-\xi)(1-\eta^2)$	

Appendix H
Analytical Evaluation of Element Matrices and Vectors

The element matrices and vectors appearing in (8.104) for the discretized ADE will be analytically evaluated for certain types of elements, where we assume constant coefficients within each element. A necessary and sufficient condition for the analytical evaluation is that the Jacobian is *constant*. This is always given for the linear 1D element, the linear 2D triangle and the linear 3D tetrahedron, for which the Jacobians are independent of the element shapes. However, to evaluate analytically quadrilateral, hexahedral, pentahedral or pyramidal elements, geometric simplifications are necessary to attain constant Jacobians for undisturbed elements. These can be shown for a quadrilateral element if simplifying to a rectangle or parallelogram, for a hexahedral element if simplifying to a brick or parallelepiped, for a pentahedral element if simplifying to a triangular prism with parallel top and bottom surfaces and for a pyramidal element if simplifying to a pyramid with parallelogram or rectangular base and oblique shape.

H.1 Linear 1D Element

We consider the linear 2-node element e as shown in Fig. H.1 with the shape functions at the nodes 1 and 2 (cf. Appendix G, Table G.1a)

$$\begin{aligned} N_1^e &= \tfrac{1}{2}(1-\xi) \\ N_2^e &= \tfrac{1}{2}(1+\xi) \end{aligned} \tag{H.1}$$

and their derivatives

$$\begin{aligned} \frac{\partial N_1^e}{\partial \xi} &= -\frac{1}{2} \\ \frac{\partial N_2^e}{\partial \xi} &= \frac{1}{2} \end{aligned} \tag{H.2}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

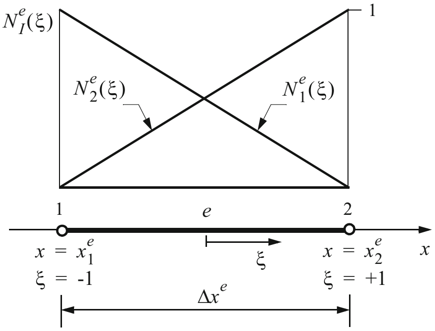


Fig. H.1 Basis functions N_I^e ($I = 1, 2$) of the linear 1D element

for $(-1 \leq \xi \leq 1)$. Furthermore, we have for the element, cf. (8.71)

$$x = N_1^e x_1^e + N_2^e x_2^e \tag{H.3}$$

and with (8.117) and (8.120) we obtain the Jacobian (and its determinant)

$$J_{11}^e = |\boldsymbol{J}^e| = \frac{\partial x}{\partial \xi} = \frac{\partial N_1^e}{\partial \xi} x_1^e + \frac{\partial N_2^e}{\partial \xi} x_2^e = \frac{\Delta x^e}{2} \tag{H.4}$$

which is constant, and the inverse Jacobian according to (8.119)

$$(\boldsymbol{J}^e)^{-1} = \frac{1}{|\boldsymbol{J}^e|} = \frac{2}{\Delta x^e} \tag{H.5}$$

where Δx^e is the element length (Fig. H.1). Then, the divergence term (8.118) becomes with (H.2)

$$\begin{pmatrix} \nabla N_1^e \\ \nabla N_2^e \end{pmatrix} = (\boldsymbol{J}^e)^{-1} \cdot \begin{pmatrix} \frac{\partial N_1^e}{\partial \xi} \\ \frac{\partial N_2^e}{\partial \xi} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\Delta x^e} \\ \frac{1}{\Delta x^e} \end{pmatrix} \tag{H.6}$$

Using (8.122) it is

$$d\Omega^e = dx = |\boldsymbol{J}^e| d\xi = \tfrac{\Delta x^e}{2} d\xi \tag{H.7}$$

Thus, the matrices and vectors of the ADE convective form (8.104) to (8.105) become for element e:

$$\begin{aligned} \boldsymbol{O}^e &= \int_{\Omega^e} \acute{\mathcal{R}}^e \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e \\ N_2^e N_1^e & N_2^e N_2^e \end{pmatrix} d\Omega^e \\ &= \tfrac{\acute{\mathcal{R}}^e}{4} \int_{-1}^{1} \begin{pmatrix} (1-\xi)^2 & 1-\xi^2 \\ 1-\xi^2 & (1+\xi)^2 \end{pmatrix} \tfrac{\Delta x^e}{2} d\xi = \tfrac{\acute{\mathcal{R}}^e \Delta x^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{A}^e &= \int_{\Omega^e} \begin{pmatrix} N_1^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_1^e(\boldsymbol{q}^e \cdot \nabla N_2^e) \\ N_2^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_2^e(\boldsymbol{q}^e \cdot \nabla N_2^e) \end{pmatrix} d\Omega^e \\
&= \tfrac{q^e}{2\Delta x^e} \int_{-1}^{1} \begin{pmatrix} -(1-\xi) & 1-\xi \\ -(1+\xi) & 1+\xi \end{pmatrix} \tfrac{\Delta x^e}{2} d\xi = \tfrac{q^e}{2} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \\
\boldsymbol{C}^e &= \int_{\Omega^e} \begin{pmatrix} \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) \\ \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) \end{pmatrix} d\Omega^e \\
&= \tfrac{D^e}{(\Delta x^e)^2} \int_{-1}^{1} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tfrac{\Delta x^e}{2} d\xi = \tfrac{D^e}{\Delta x^e} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \\
\boldsymbol{R}^e &= \int_{\Omega^e} (\vartheta^e + Q^e) \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e \\ N_2^e N_1^e & N_2^e N_2^e \end{pmatrix} d\Omega^e \\
&= \tfrac{\vartheta^e + Q^e}{4} \int_{-1}^{1} \begin{pmatrix} (1-\xi)^2 & 1-\xi^2 \\ 1-\xi^2 & (1+\xi)^2 \end{pmatrix} \tfrac{\Delta x^e}{2} d\xi = \tfrac{(\vartheta^e + Q^e)\Delta x^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{B}^e &= \int_{\Gamma_C^e} \Phi^e \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e \\ N_2^e N_1^e & N_2^e N_2^e \end{pmatrix} d\Gamma^e - \int_{\Gamma_{N_O}^e} \begin{pmatrix} N_1^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_1^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} \\ N_2^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_2^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} \end{pmatrix} d\Gamma^e \\
&= \tfrac{\Phi^e}{4} \begin{pmatrix} (1-\xi)^2 & 1-\xi^2 \\ 1-\xi^2 & (1+\xi)^2 \end{pmatrix}_{\xi=\xi_2=1}^{\xi=\xi_1=-1} - \tfrac{D^e}{2\Delta x^e} \begin{pmatrix} -(1-\xi) & 1-\xi \\ -(1+\xi) & 1+\xi \end{pmatrix}_{\xi=\xi_2=1}^{\xi=\xi_1=-1} \\
&= \Phi^e \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \tfrac{D^e}{\Delta x^e} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}_{\Gamma_{N_O}^e}
\end{aligned}
\tag{H.8}
$$

$$
\begin{aligned}
\boldsymbol{H}^e &= \int_{\Gamma_C^e} \Phi^e \phi_C^e \begin{pmatrix} N_1^e \\ N_2^e \end{pmatrix} d\Gamma^e - \int_{\Gamma_N^e} q_N^e \begin{pmatrix} N_1^e \\ N_2^e \end{pmatrix} d\Gamma^e \\
&= \tfrac{\Phi^e \phi_C^e}{2} \begin{pmatrix} 1-\xi \\ 1+\xi \end{pmatrix}_{\xi=\xi_2=1}^{\xi=\xi_1=-1} - \tfrac{q_N^e}{2} \begin{pmatrix} 1-\xi \\ 1+\xi \end{pmatrix}_{\xi=\xi_2=1}^{\xi=\xi_1=-1} = \Phi^e \phi_C^e \begin{pmatrix} 1 \\ 1 \end{pmatrix} - q_N^e \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
\boldsymbol{Q}^e &= \int_{\Omega^e} H^e \begin{pmatrix} N_1^e \\ N_2^e \end{pmatrix} d\Omega^e - \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t) \\
&= \tfrac{H^e}{2} \int_{-1}^{1} \begin{pmatrix} 1-\xi \\ 1+\xi \end{pmatrix} \tfrac{\Delta x^e}{2} d\xi - \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t) = \tfrac{H^e \Delta x^e}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t)
\end{aligned}
\tag{H.9}
$$

where for convenience we assume constant parameters (storage coefficient $\acute{\mathcal{R}}^e$, flux $\boldsymbol{q}^e$, dispersion $\boldsymbol{D}^e$, decay rate ϑ^e, flow supply Q^e, transfer coefficient Φ^e and source/sink H^e) within the element. Finally, the spatially discretized ADE in the convective form (8.99) can be summarized as

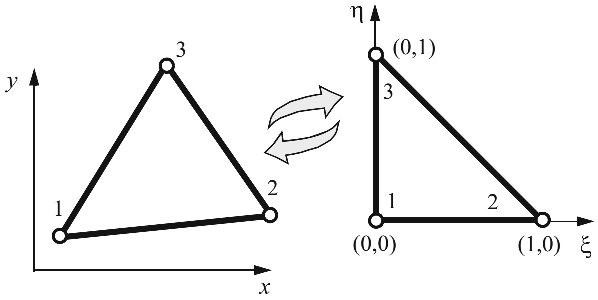


Fig. H.2 Linear 3-node triangle in the global and local coordinate system

$$\sum_e \left\{ \frac{\acute{\mathcal{R}}^e \Delta x^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} \frac{d\phi_1^e}{dt} \\ \frac{d\phi_2^e}{dt} \end{pmatrix} + \left[\frac{q^e}{2} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} + \frac{D^e}{\Delta x^e} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{(\vartheta^e + Q^e)\Delta x^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \right.\right.$$
$$\left.\left. + \Phi^e \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{D^e}{\Delta x^e} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}_{\Gamma_{N_O}} \right] \cdot \begin{pmatrix} \phi_1^e \\ \phi_2^e \end{pmatrix} - \Phi^e \phi_C^e \begin{pmatrix} 1 \\ 1 \end{pmatrix} + q_N^e \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{H^e \Delta x^e}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$
$$+ \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t) = \mathbf{0} \qquad \text{(H.10)}$$

Similar expressions can be obtained for the divergence form of ADE (8.98).

H.2 Linear 2D Triangle

We consider the linear 3-node triangular element e as shown in Fig. H.2 (cf. Appendix G, Table G.2a) having the following shape functions

$$\begin{aligned} N_1^e &= 1 - \xi - \eta \\ N_2^e &= \xi \\ N_3^e &= \eta \end{aligned} \qquad \text{(H.11)}$$

and local derivatives

$$\begin{aligned} \frac{\partial N_1^e}{\partial \xi} &= -1, & \frac{\partial N_1^e}{\partial \eta} &= -1 \\ \frac{\partial N_2^e}{\partial \xi} &= 1, & \frac{\partial N_2^e}{\partial \eta} &= 0 \\ \frac{\partial N_3^e}{\partial \xi} &= 0, & \frac{\partial N_3^e}{\partial \eta} &= 1 \end{aligned} \qquad \text{(H.12)}$$

for $(0 \leq \xi, \eta \leq 1)$. Furthermore, we have for the triangular element, cf. (8.71)

$$\begin{aligned} x &= N_1^e x_1^e + N_2^e x_2^e + N_3^e x_3^e \\ y &= N_1^e y_1^e + N_2^e y_2^e + N_3^e y_3^e \end{aligned} \qquad \text{(H.13)}$$

where x_I^e, y_I^e $(I = 1, 2, 3)$ correspond to the Cartesian coordinates of the vertices (nodes) of the triangle. Using (H.12) and (H.13) in (8.116), (8.119), and (8.120) we obtain the Jacobian

$$\boldsymbol{J}^e = \begin{pmatrix} x_2^e - x_1^e & y_2^e - y_1^e \\ x_3^e - x_1^e & y_3^e - y_1^e \end{pmatrix}, \tag{H.14}$$

the inverse Jacobian

$$(\boldsymbol{J}^e)^{-1} = \frac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} y_3^e - y_1^e & y_1^e - y_2^e \\ x_1^e - x_3^e & x_2^e - x_1^e \end{pmatrix} \tag{H.15}$$

and its determinant

$$|\boldsymbol{J}^e| = x_1^e(y_2^e - y_3^e) + x_2^e(y_3^e - y_1^e) + x_3^e(y_1^e - y_2^e) = 2A^e \tag{H.16}$$

which is twice the area A^e of the triangle. We recognize that the triangle has the advantage of having a constant Jacobian (and its inverse) independent of the element shape in the Cartesian coordinates (we shall see further below this is not the case for the quadrilateral). The global derivatives (8.118) result with (H.12) in

$$\begin{aligned}
\nabla N_1^e &= \tfrac{1}{2A^e} \begin{pmatrix} y_3^e - y_1^e & y_1^e - y_2^e \\ x_1^e - x_3^e & x_2^e - x_1^e \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial N_1^e}{\partial \xi} \\ \frac{\partial N_1^e}{\partial \eta} \end{pmatrix} = \tfrac{1}{2A^e} \begin{pmatrix} y_2^e - y_3^e \\ x_3^e - x_2^e \end{pmatrix} = \tfrac{1}{2A^e} \underbrace{\begin{pmatrix} S_{11}^e \\ S_{21}^e \end{pmatrix}}_{\boldsymbol{S}^e \cdot \boldsymbol{e}_1} \\
\nabla N_2^e &= \tfrac{1}{2A^e} \begin{pmatrix} y_3^e - y_1^e & y_1^e - y_2^e \\ x_1^e - x_3^e & x_2^e - x_1^e \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial N_2^e}{\partial \xi} \\ \frac{\partial N_2^e}{\partial \eta} \end{pmatrix} = \tfrac{1}{2A^e} \begin{pmatrix} y_3^e - y_1^e \\ x_1^e - x_3^e \end{pmatrix} = \tfrac{1}{2A^e} \underbrace{\begin{pmatrix} S_{12}^e \\ S_{22}^e \end{pmatrix}}_{\boldsymbol{S}^e \cdot \boldsymbol{e}_2} \\
\nabla N_3^e &= \tfrac{1}{2A^e} \begin{pmatrix} y_3^e - y_1^e & y_1^e - y_2^e \\ x_1^e - x_3^e & x_2^e - x_1^e \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial N_3^e}{\partial \xi} \\ \frac{\partial N_3^e}{\partial \eta} \end{pmatrix} = \tfrac{1}{2A^e} \begin{pmatrix} y_1^e - y_2^e \\ x_2^e - x_1^e \end{pmatrix} = \tfrac{1}{2A^e} \underbrace{\begin{pmatrix} S_{13}^e \\ S_{23}^e \end{pmatrix}}_{\boldsymbol{S}^e \cdot \boldsymbol{e}_3}
\end{aligned} \tag{H.17}$$

introducing the coordinate matrix

$$\boldsymbol{S}^e = \begin{pmatrix} S_{11}^e & S_{12}^e & S_{13}^e \\ S_{21}^e & S_{22}^e & S_{23}^e \end{pmatrix} = \begin{pmatrix} y_2^e - y_3^e & y_3^e - y_1^e & y_1^e - y_2^e \\ x_3^e - x_2^e & x_1^e - x_3^e & x_2^e - x_1^e \end{pmatrix} \tag{H.18}$$

where $\boldsymbol{e}_I$ are nodal base vectors defined as

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \boldsymbol{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{H.19}$$

With (8.122) and (H.16) as well as (8.123) and (H.15) it is

$$d\Omega^e = |\boldsymbol{J}^e| d\xi d\eta = 2A^e d\xi d\eta \tag{H.20}$$

and

$$d\Gamma^e = \begin{cases} l_{12}^e d\xi & \text{on edge } \widehat{12},\ \eta = 0 \\ l_{13}^e d\eta & \text{on edge } \widehat{13},\ \xi = 0 \\ l_{23}^e d\xi & \text{on edge } \widehat{23},\ \eta = 1 - \xi \end{cases} \tag{H.21}$$

where

$$\begin{aligned} l_{12}^e &= \left\| \begin{pmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \end{pmatrix} \right\| &&= \left\| \begin{pmatrix} x_2^e - x_1^e \\ y_2^e - y_1^e \end{pmatrix} \right\| = \sqrt{(x_2^e - x_1^e)^2 + (y_2^e - y_1^e)^2} \\ l_{13}^e &= \left\| \begin{pmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \end{pmatrix} \right\| &&= \left\| \begin{pmatrix} x_3^e - x_1^e \\ y_3^e - y_1^e \end{pmatrix} \right\| = \sqrt{(x_3^e - x_1^e)^2 + (y_3^e - y_1^e)^2} \\ l_{23}^e &= \left\| \begin{pmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \end{pmatrix} - \begin{pmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \end{pmatrix} \right\| &&= \left\| \begin{pmatrix} x_3^e - x_2^e \\ y_3^e - y_2^e \end{pmatrix} \right\| = \sqrt{(x_3^e - x_2^e)^2 + (y_3^e - y_2^e)^2} \end{aligned} \tag{H.22}$$

Thus, the matrices and vectors of the ADE convective form (8.104) and (8.105) become for element e:

$$\begin{aligned} \boldsymbol{O}^e &= \int_{\Omega^e} \acute{\mathcal{R}}^e \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e & N_1^e N_3^e \\ N_2^e N_1^e & N_2^e N_2^e & N_2^e N_3^e \\ N_3^e N_1^e & N_3^e N_2^e & N_3^e N_3^e \end{pmatrix} d\Omega^e \\ &= \acute{\mathcal{R}}^e \int_0^1 \int_0^{1-\xi} \begin{pmatrix} (1-\xi-\eta)^2 & (1-\xi-\eta)\xi & (1-\xi-\eta)\eta \\ (1-\xi-\eta)\xi & \xi^2 & \xi\eta \\ (1-\xi-\eta)\eta & \xi\eta & \eta^2 \end{pmatrix} 2A^e d\eta d\xi = \frac{\acute{\mathcal{R}}^e A^e}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \\ \boldsymbol{A}^e &= \int_{\Omega^e} \begin{pmatrix} N_1^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_1^e(\boldsymbol{q}^e \cdot \nabla N_2^e) & N_1^e(\boldsymbol{q}^e \cdot \nabla N_3^e) \\ N_2^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_2^e(\boldsymbol{q}^e \cdot \nabla N_2^e) & N_2^e(\boldsymbol{q}^e \cdot \nabla N_3^e) \\ N_3^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_3^e(\boldsymbol{q}^e \cdot \nabla N_2^e) & N_3^e(\boldsymbol{q}^e \cdot \nabla N_3^e) \end{pmatrix} d\Omega^e \\ &= \frac{q^e}{2A^e} \int_0^1 \int_0^{1-\xi} \begin{pmatrix} (1-\xi-\eta)(S_{11}^e + S_{21}^e) & (1-\xi-\eta)(S_{12}^e + S_{22}^e) & (1-\xi-\eta)(S_{13}^e + S_{23}^e) \\ \xi(S_{11}^e + S_{21}^e) & \xi(S_{12}^e + S_{22}^e) & \xi(S_{13}^e + S_{23}^e) \\ \eta(S_{11}^e + S_{21}^e) & \eta(S_{12}^e + S_{22}^e) & \eta(S_{13}^e + S_{23}^e) \end{pmatrix} 2A^e d\eta d\xi \\ &= \frac{q^e}{6} \begin{pmatrix} (S_{11}^e + S_{21}^e) & (S_{12}^e + S_{22}^e) & (S_{13}^e + S_{23}^e) \\ (S_{11}^e + S_{21}^e) & (S_{12}^e + S_{22}^e) & (S_{13}^e + S_{23}^e) \\ (S_{11}^e + S_{21}^e) & (S_{12}^e + S_{22}^e) & (S_{13}^e + S_{23}^e) \end{pmatrix} \\ \boldsymbol{C}^e &= \int_{\Omega^e} \begin{pmatrix} \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) & \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_3^e) \\ \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) & \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_3^e) \\ \nabla N_3^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_3^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) & \nabla N_3^e \cdot (\boldsymbol{D}^e \cdot \nabla N_3^e) \end{pmatrix} d\Omega^e \\ &= \frac{D^e}{(2A^e)^2} \int_0^1 \int_0^{1-\xi} \begin{pmatrix} ((S_{11}^e)^2 + (S_{21}^e)^2) & (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e) & (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e) \\ (S_{12}^e S_{11}^e + S_{22}^e S_{21}^e) & ((S_{12}^e)^2 + (S_{22}^e)^2) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e) \\ (S_{13}^e S_{11}^e + S_{23}^e S_{21}^e) & (S_{13}^e S_{12}^e + S_{23}^e S_{22}^e) & ((S_{13}^e)^2 + (S_{23}^e)^2) \end{pmatrix} 2A^e d\eta d\xi \end{aligned}$$

$$= \tfrac{D^e}{4A^e} \begin{pmatrix} ((S_{11}^e)^2 + (S_{21}^e)^2) & (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e) & (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e) \\ (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e) & ((S_{12}^e)^2 + (S_{22}^e)^2) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e) \\ (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e) & ((S_{13}^e)^2 + (S_{23}^e)^2) \end{pmatrix}$$

$$\boldsymbol{R}^e = \textstyle\int_{\Omega^e} (\vartheta^e + Q^e) \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e & N_1^e N_3^e \\ N_2^e N_1^e & N_2^e N_2^e & N_2^e N_3^e \\ N_3^e N_1^e & N_3^e N_2^e & N_3^e N_3^e \end{pmatrix} d\Omega^e$$

$$= (\vartheta^e + Q^e) \textstyle\int_0^1 \int_0^{1-\xi} \begin{pmatrix} (1-\xi-\eta)^2 & (1-\xi-\eta)\xi & (1-\xi-\eta)\eta \\ (1-\xi-\eta)\xi & \xi^2 & \xi\eta \\ (1-\xi-\eta)\eta & \xi\eta & \eta^2 \end{pmatrix} 2A^e d\eta d\xi$$

$$= \frac{(\vartheta^e + Q^e)A^e}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \tag{H.23}$$

$$\boldsymbol{B}^e = \textstyle\int_{\Gamma_C^e} \Phi^e \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e & N_1^e N_3^e \\ N_2^e N_1^e & N_2^e N_2^e & N_2^e N_3^e \\ N_3^e N_1^e & N_3^e N_2^e & N_3^e N_3^e \end{pmatrix} d\Gamma^e$$

$$- \textstyle\int_{\Gamma_{N_O}^e} \begin{pmatrix} N_1^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_1^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} & N_1^e(\boldsymbol{D} \cdot \nabla N_3^e) \cdot \boldsymbol{n} \\ N_2^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_2^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} & N_2^e(\boldsymbol{D} \cdot \nabla N_3^e) \cdot \boldsymbol{n} \\ N_3^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_3^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} & N_3^e(\boldsymbol{D} \cdot \nabla N_3^e) \cdot \boldsymbol{n} \end{pmatrix} d\Gamma^e$$

$$= \Phi^e \int_0^1 \begin{pmatrix} (1-\xi-\eta)^2 & (1-\xi-\eta)\xi & (1-\xi-\eta)\eta \\ (1-\xi-\eta)\xi & \xi^2 & \xi\eta \\ (1-\xi-\eta)\eta & \xi\eta & \eta^2 \end{pmatrix} \sqrt{\ldots}\,(d\xi, d\eta)$$

$$- \tfrac{D^e}{2A^e} \int_0^1 \begin{pmatrix} (1-\xi-\eta)(S_{11}^e n_1 + S_{21}^e n_2) & (1-\xi-\eta)(S_{12}^e n_1 + S_{22}^e n_2) & (1-\xi-\eta)(S_{13}^e n_1 + S_{23}^e n_2) \\ \xi(S_{11}^e n_1 + S_{21}^e n_2) & \xi(S_{12}^e n_1 + S_{22}^e n_2) & \xi(S_{13}^e n_1 + S_{23}^e n_2) \\ \eta(S_{11}^e n_1 + S_{21}^e n_2) & \eta(S_{12}^e n_1 + S_{22}^e n_2) & \eta(S_{13}^e n_1 + S_{23}^e n_2) \end{pmatrix}$$

$$\times \sqrt{\ldots}\,(d\xi, d\eta)$$

$$= \begin{cases} \tfrac{\Phi^e}{6} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} l_{12}^e - & \\ \tfrac{D^e}{4A^e} \begin{pmatrix} (S_{11}^e n_1 + S_{21}^e n_2) & (S_{12}^e n_1 + S_{22}^e n_2) & (S_{13}^e n_1 + S_{23}^e n_2) \\ (S_{11}^e n_1 + S_{21}^e n_2) & (S_{12}^e n_1 + S_{22}^e n_2) & (S_{13}^e n_1 + S_{23}^e n_2) \\ 0 & 0 & 0 \end{pmatrix}_{\Gamma_{N_O}} l_{12}^e + & \text{on edge } \widehat{12} \\ \tfrac{\Phi^e}{6} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} l_{13}^e - & \\ \tfrac{D^e}{4A^e} \begin{pmatrix} (S_{11}^e n_1 + S_{21}^e n_2) & (S_{12}^e n_1 + S_{22}^e n_2) & (S_{13}^e n_1 + S_{23}^e n_2) \\ 0 & 0 & 0 \\ (S_{11}^e n_1 + S_{21}^e n_2) & (S_{12}^e n_1 + S_{22}^e n_2) & (S_{13}^e n_1 + S_{23}^e n_2) \end{pmatrix}_{\Gamma_{N_O}} l_{13}^e + & \text{on edge } \widehat{13} \\ \tfrac{\Phi^e}{6} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} l_{23}^e - & \\ \tfrac{D^e}{4A^e} \begin{pmatrix} 0 & 0 & 0 \\ (S_{11}^e n_1 + S_{21}^e n_2) & (S_{12}^e n_1 + S_{22}^e n_2) & (S_{13}^e n_1 + S_{23}^e n_2) \\ (S_{11}^e n_1 + S_{21}^e n_2) & (S_{12}^e n_1 + S_{22}^e n_2) & (S_{13}^e n_1 + S_{23}^e n_2) \end{pmatrix}_{\Gamma_{N_O}} l_{23}^e & \text{on edge } \widehat{23} \end{cases}$$

$$\begin{aligned}
\boldsymbol{H}^e &= \int_{\Gamma_C^e} \varPhi^e \phi_C^e \begin{pmatrix} N_1^e \\ N_2^e \\ N_3^e \end{pmatrix} d\Gamma^e - \int_{\Gamma_N^e} q_N^e \begin{pmatrix} N_1^e \\ N_2^e \\ N_3^e \end{pmatrix} d\Gamma^e \\
&= \varPhi^e \phi_C^e \int_0^1 \begin{pmatrix} 1-\xi-\eta \\ \xi \\ \eta \end{pmatrix} \sqrt{\ldots}\,(d\xi, d\eta) - q_N^e \int_0^1 \begin{pmatrix} 1-\xi-\eta \\ \xi \\ \eta \end{pmatrix} \sqrt{\ldots}\,(d\xi, d\eta) \\
&= \begin{cases}
\left[\frac{\varPhi^e \phi_C^e}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - \frac{q_N^e}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right] l_{12}^e + & \text{on edge } \widehat{12} \\
\left[\frac{\varPhi^e \phi_C^e}{2} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - \frac{q_N^e}{2} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right] l_{13}^e + & \text{on edge } \widehat{13} \\
\left[\frac{\varPhi^e \phi_C^e}{2} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} - \frac{q_N^e}{2} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right] l_{23}^e & \text{on edge } \widehat{23}
\end{cases}
\end{aligned} \qquad \text{(H.24)}$$

$$\begin{aligned}
\boldsymbol{Q}^e &= \int_{\Omega^e} H^e \begin{pmatrix} N_1^e \\ N_2^e \\ N_3^e \end{pmatrix} d\Omega^e - \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t) \\
&= H^e \int_0^1 \int_0^{1-\xi} \begin{pmatrix} 1-\xi-\eta \\ \xi \\ \eta \end{pmatrix} 2A^e d\eta d\xi - \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t) \\
&= \frac{H^e A^e}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t)
\end{aligned} \qquad \text{(H.25)}$$

where for convenience we assume constant parameters (storage coefficient $\acute{\mathcal{R}}^e$, flux $\boldsymbol{q}^e$, dispersion $\boldsymbol{D}^e$, decay rate ϑ^e, flow supply Q^e, transfer coefficient $\varPhi^e$ and source/sink H^e) within the element. In (H.24) the components of the unit normal vector $\boldsymbol{n}$ in the Cartesian $x-$ and $y-$coordinate direction on the outflow boundary Γ_{N_O} are indicated by n_1 and n_2, respectively. Finally, the spatially discretized ADE in the convective form (8.99) can be summarized as

$$\begin{aligned}
&\sum_e \left\{ \frac{\acute{\mathcal{R}}^e A^e}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} \frac{d\phi_1^e}{dt} \\ \frac{d\phi_2^e}{dt} \\ \frac{d\phi_3^e}{dt} \end{pmatrix} + \left[\frac{q^e}{6} \begin{pmatrix} (S_{11}^e + S_{21}^e) & (S_{12}^e + S_{22}^e) & (S_{13}^e + S_{23}^e) \\ (S_{11}^e + S_{21}^e) & (S_{12}^e + S_{22}^e) & (S_{13}^e + S_{23}^e) \\ (S_{11}^e + S_{21}^e) & (S_{12}^e + S_{22}^e) & (S_{13}^e + S_{23}^e) \end{pmatrix} \right.\right. \\
&+ \frac{D^e}{4A^e} \begin{pmatrix} ((S_{11}^e)^2 + (S_{21}^e)^2) & (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e) & (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e) \\ (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e) & ((S_{12}^e)^2 + (S_{22}^e)^2) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e) \\ (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e) & ((S_{13}^e)^2 + (S_{23}^e)^2) \end{pmatrix} \\
&+ \frac{(\vartheta^e + Q^e) A^e}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} + \frac{\varPhi^e}{6} \left(\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} l_{12}^e + \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} l_{13}^e + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} l_{23}^e \right)
\end{aligned}$$

$$
\begin{aligned}
&-\frac{D^e}{4A^e}\Bigg(\begin{pmatrix} (S^e_{11}n_1 + S^e_{21}n_2) & (S^e_{12}n_1 + S^e_{22}n_2) & (S^e_{13}n_1 + S^e_{23}n_2) \\ (S^e_{11}n_1 + S^e_{21}n_2) & (S^e_{12}n_1 + S^e_{22}n_2) & (S^e_{13}n_1 + S^e_{23}n_2) \\ 0 & 0 & 0 \end{pmatrix} l^e_{12} \\
&\qquad + \begin{pmatrix} (S^e_{11}n_1 + S^e_{21}n_2) & (S^e_{12}n_1 + S^e_{22}n_2) & (S^e_{13}n_1 + S^e_{23}n_2) \\ 0 & 0 & 0 \\ (S^e_{11}n_1 + S^e_{21}n_2) & (S^e_{12}n_1 + S^e_{22}n_2) & (S^e_{13}n_1 + S^e_{23}n_2) \end{pmatrix} l^e_{13} \\
&\qquad + \begin{pmatrix} 0 & 0 & 0 \\ (S^e_{11}n_1 + S^e_{21}n_2) & (S^e_{12}n_1 + S^e_{22}n_2) & (S^e_{13}n_1 + S^e_{23}n_2) \\ (S^e_{11}n_1 + S^e_{21}n_2) & (S^e_{12}n_1 + S^e_{22}n_2) & (S^e_{13}n_1 + S^e_{23}n_2) \end{pmatrix} l^e_{23}\Bigg)_{\Gamma_{NO}} \\
&\Bigg] \cdot \begin{pmatrix} \phi^e_1 \\ \phi^e_2 \\ \phi^e_3 \end{pmatrix} - \frac{\Phi^e \phi^e_C - q^e_N}{2}\Bigg[\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} l^e_{12} + \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} l^e_{13} + \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} l^e_{23}\Bigg] - \frac{H^e A^e}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\Bigg\} \\
&+ \sum\nolimits_w \phi_w(\boldsymbol{x}_w) Q_w(t) = \mathbf{0}
\end{aligned}
\tag{H.26}
$$

Similar expressions can be obtained for the divergence form of ADE (8.98).

H.3 Linear 3D Tetrahedron

We consider the linear 4-node tetrahedral element e as shown in Fig. H.3 (cf. Appendix G, Table G.3a) having the following shape functions

$$
\begin{aligned}
N^e_1 &= 1 - \xi - \eta - \zeta \\
N^e_2 &= \xi \\
N^e_3 &= \eta \\
N^e_4 &= \zeta
\end{aligned}
\tag{H.27}
$$

and local derivatives

$$
\begin{array}{lll}
\frac{\partial N^e_1}{\partial \xi} = -1, & \frac{\partial N^e_1}{\partial \eta} = -1 & \frac{\partial N^e_1}{\partial \zeta} = -1 \\
\frac{\partial N^e_2}{\partial \xi} = 1, & \frac{\partial N^e_2}{\partial \eta} = 0 & \frac{\partial N^e_2}{\partial \zeta} = 0 \\
\frac{\partial N^e_3}{\partial \xi} = 0, & \frac{\partial N^e_3}{\partial \eta} = 1 & \frac{\partial N^e_3}{\partial \zeta} = 0 \\
\frac{\partial N^e_4}{\partial \xi} = 0, & \frac{\partial N^e_4}{\partial \eta} = 0 & \frac{\partial N^e_4}{\partial \zeta} = 1
\end{array}
\tag{H.28}
$$

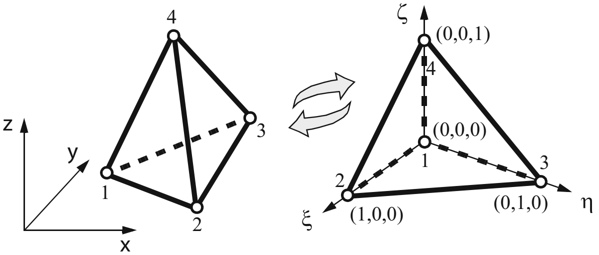


Fig. H.3 Linear 4-node tetrahedron in the global and local coordinate system

for $(0 \leq \xi, \eta, \zeta \leq 1)$. It is for the tetrahedral element, cf. (8.71)

$$\begin{aligned} x &= N_1^e x_1^e + N_2^e x_2^e + N_3^e x_3^e + N_4^e x_4^e \\ y &= N_1^e y_1^e + N_2^e y_2^e + N_3^e y_3^e + N_4^e y_4^e \\ z &= N_1^e z_1^e + N_2^e z_2^e + N_3^e z_3^e + N_4^e z_4^e \end{aligned} \tag{H.29}$$

where x_I^e, y_I^e, z_I^e $(I = 1, 2, 3, 4)$ correspond to the Cartesian coordinates of the vertices (nodes) of the tetrahedron.Using (H.28) and (H.29) in (8.115) and (8.119) we obtain the Jacobian

$$\boldsymbol{J}^e = \begin{pmatrix} x_2^e - x_1^e & y_2^e - y_1^e & z_2^e - z_1^e \\ x_3^e - x_1^e & y_3^e - y_1^e & z_3^e - z_1^e \\ x_4^e - x_1^e & y_4^e - y_1^e & z_4^e - z_1^e \end{pmatrix}, \tag{H.30}$$

the inverse Jacobian

$$(\boldsymbol{J}^e)^{-1} = \frac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} A_{11}^e & A_{12}^e & A_{13}^e \\ A_{21}^e & A_{22}^e & A_{23}^e \\ A_{31}^e & A_{32}^e & A_{33}^e \end{pmatrix} \tag{H.31}$$

where

$$\begin{aligned} A_{11}^e &= z_1^e(y_4^e - y_3^e) + z_3^e(y_1^e - y_4^e) + z_4^e(y_3^e - y_1^e) \\ A_{12}^e &= z_1^e(y_2^e - y_4^e) + z_2^e(y_4^e - y_1^e) + z_4^e(y_1^e - y_2^e) \\ A_{13}^e &= z_1^e(y_3^e - y_2^e) + z_2^e(y_1^e - y_3^e) + z_3^e(y_2^e - y_1^e) \\ A_{21}^e &= z_1^e(x_3^e - x_4^e) + z_3^e(x_4^e - x_1^e) + z_4^e(x_1^e - x_3^e) \\ A_{22}^e &= z_1^e(x_4^e - x_2^e) + z_2^e(x_1^e - x_4^e) + z_4^e(x_2^e - x_1^e) \\ A_{23}^e &= z_1^e(x_2^e - x_3^e) + z_2^e(x_3^e - x_1^e) + z_3^e(x_1^e - x_2^e) \\ A_{31}^e &= y_1^e(x_4^e - x_3^e) + y_3^e(x_1^e - x_4^e) + y_4^e(x_3^e - x_1^e) \\ A_{32}^e &= y_1^e(x_2^e - x_4^e) + y_2^e(x_4^e - x_1^e) + y_4^e(x_1^e - x_2^e) \\ A_{33}^e &= y_1^e(x_3^e - x_2^e) + y_2^e(x_1^e - x_3^e) + y_3^e(x_2^e - x_1^e) \end{aligned} \tag{H.32}$$

and its determinant

$$|\boldsymbol{J}^e| = (x_2^e - x_1^e)A_{11}^e + (x_3^e - x_1^e)A_{12}^e + (x_4^e - x_1^e)A_{13}^e = 6V^e \tag{H.33}$$

which is six times the volume V^e of the tetrahedron. We see that the tetrahedron has the important advantage of having a constant Jacobian (and its inverse) independent of the element shape in the Cartesian coordinates (we shall see further below this is not the case for the other 3D elements). The global derivatives (8.118) result with (H.28) in

$$\begin{aligned}
\nabla N_1^e &= \frac{1}{6V^e}\begin{pmatrix}A_{11}^e & A_{12}^e & A_{13}^e\\ A_{21}^e & A_{22}^e & A_{23}^e\\ A_{31}^e & A_{32}^e & A_{33}^e\end{pmatrix}\cdot\begin{pmatrix}\frac{\partial N_1^e}{\partial\xi}\\ \frac{\partial N_1^e}{\partial\eta}\\ \frac{\partial N_1^e}{\partial\zeta}\end{pmatrix} = \frac{1}{6V^e}\begin{pmatrix}-(A_{11}^e+A_{12}^e+A_{13}^e)\\ -(A_{21}^e+A_{22}^e+A_{23}^e)\\ -(A_{31}^e+A_{32}^e+A_{33}^e)\end{pmatrix} = \frac{1}{6V^e}\underbrace{\begin{pmatrix}S_{11}^e\\ S_{21}^e\\ S_{31}^e\end{pmatrix}}_{\boldsymbol{S}^e\cdot\boldsymbol{e}_1}\\
\nabla N_2^e &= \frac{1}{6V^e}\begin{pmatrix}A_{11}^e & A_{12}^e & A_{13}^e\\ A_{21}^e & A_{22}^e & A_{23}^e\\ A_{31}^e & A_{32}^e & A_{33}^e\end{pmatrix}\cdot\begin{pmatrix}\frac{\partial N_1^e}{\partial\xi}\\ \frac{\partial N_1^e}{\partial\eta}\\ \frac{\partial N_1^e}{\partial\zeta}\end{pmatrix} = \frac{1}{6V^e}\begin{pmatrix}A_{11}^e\\ A_{21}^e\\ A_{31}^e\end{pmatrix} = \frac{1}{6V^e}\underbrace{\begin{pmatrix}S_{12}^e\\ S_{22}^e\\ S_{32}^e\end{pmatrix}}_{\boldsymbol{S}^e\cdot\boldsymbol{e}_2}\\
\nabla N_3^e &= \frac{1}{6V^e}\begin{pmatrix}A_{11}^e & A_{12}^e & A_{13}^e\\ A_{21}^e & A_{22}^e & A_{23}^e\\ A_{31}^e & A_{32}^e & A_{33}^e\end{pmatrix}\cdot\begin{pmatrix}\frac{\partial N_1^e}{\partial\xi}\\ \frac{\partial N_1^e}{\partial\eta}\\ \frac{\partial N_1^e}{\partial\zeta}\end{pmatrix} = \frac{1}{6V^e}\begin{pmatrix}A_{12}^e\\ A_{22}^e\\ A_{32}^e\end{pmatrix} = \frac{1}{6V^e}\underbrace{\begin{pmatrix}S_{13}^e\\ S_{23}^e\\ S_{33}^e\end{pmatrix}}_{\boldsymbol{S}^e\cdot\boldsymbol{e}_3}\\
\nabla N_4^e &= \frac{1}{6V^e}\begin{pmatrix}A_{11}^e & A_{12}^e & A_{13}^e\\ A_{21}^e & A_{22}^e & A_{23}^e\\ A_{31}^e & A_{32}^e & A_{33}^e\end{pmatrix}\cdot\begin{pmatrix}\frac{\partial N_1^e}{\partial\xi}\\ \frac{\partial N_1^e}{\partial\eta}\\ \frac{\partial N_1^e}{\partial\zeta}\end{pmatrix} = \frac{1}{6V^e}\begin{pmatrix}A_{13}^e\\ A_{23}^e\\ A_{33}^e\end{pmatrix} = \frac{1}{6V^e}\underbrace{\begin{pmatrix}S_{14}^e\\ S_{24}^e\\ S_{34}^e\end{pmatrix}}_{\boldsymbol{S}^e\cdot\boldsymbol{e}_4}
\end{aligned}\tag{H.34}$$

introducing the coordinate matrix

$$\boldsymbol{S}^e = \begin{pmatrix}S_{11}^e & S_{12}^e & S_{13}^e & S_{14}^e\\ S_{21}^e & S_{22}^e & S_{23}^e & S_{24}^e\\ S_{31}^e & S_{32}^e & S_{33}^e & S_{34}^e\end{pmatrix} = \begin{pmatrix}A_{10}^e & A_{11}^e & A_{12}^e & A_{13}^e\\ A_{20}^e & A_{21}^e & A_{22}^e & A_{23}^e\\ A_{30}^e & A_{31}^e & A_{32}^e & A_{33}^e\end{pmatrix} \tag{H.35}$$

with

$$\begin{aligned}
A_{10}^e &= z_2^e(y_3^e - y_4^e) + z_3^e(y_4^e - y_2^e) + z_4^e(y_2^e - y_3^e)\\
A_{20}^e &= z_2^e(x_4^e - x_3^e) + z_3^e(x_2^e - x_4^e) + z_4^e(x_3^e - x_2^e)\\
A_{30}^e &= y_2^e(x_3^e - x_4^e) + y_3^e(x_4^e - x_2^e) + y_4^e(x_2^e - x_3^e)
\end{aligned}\tag{H.36}$$

where $\boldsymbol{e}_I$ are nodal base vectors defined as

$$\boldsymbol{e}_1 = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}\quad \boldsymbol{e}_2 = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}\quad \boldsymbol{e}_3 = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}\quad \boldsymbol{e}_4 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \tag{H.37}$$

With (8.122) and (H.33) as well as (8.123) and (H.31) it is

$$d\Omega^e = |\boldsymbol{J}^e| d\xi d\eta d\zeta = 6V^e d\xi d\eta d\zeta \tag{H.38}$$

and

$$d\Gamma^e = \begin{cases} 2\, A^e_{\widehat{123}} d\eta d\xi & \text{on area } \widehat{123},\ \eta = 1-\xi,\ \zeta = 0 \\ 2\, A^e_{\widehat{124}} d\zeta d\xi & \text{on area } \widehat{124},\ \zeta = 1-\xi,\ \eta = 0 \\ 2\, A^e_{\widehat{134}} d\zeta d\eta & \text{on area } \widehat{134},\ \zeta = 1-\eta,\ \xi = 0 \\ 2\, A^e_{\widehat{234}} d\zeta d\xi & \text{on area } \widehat{234},\ \eta = 1-\xi-\zeta \end{cases} \tag{H.39}$$

where

$$\begin{aligned}
2\, A^e_{\widehat{123}} &= \left\| \begin{pmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \\ \frac{\partial z}{\partial \xi} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \\ \frac{\partial z}{\partial \eta} \end{pmatrix} \right\| = \left\| \begin{pmatrix} A^e_{13} \\ A^e_{23} \\ A^e_{33} \end{pmatrix} \right\| = \sqrt{(A^e_{13})^2 + (A^e_{23})^2 + (A^e_{33})^2} \\
2\, A^e_{\widehat{124}} &= \left\| \begin{pmatrix} \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \zeta} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \\ \frac{\partial z}{\partial \xi} \end{pmatrix} \right\| = \left\| \begin{pmatrix} A^e_{12} \\ A^e_{22} \\ A^e_{32} \end{pmatrix} \right\| = \sqrt{(A^e_{12})^2 + (A^e_{22})^2 + (A^e_{32})^2} \\
2\, A^e_{\widehat{134}} &= \left\| \begin{pmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \\ \frac{\partial z}{\partial \eta} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \zeta} \end{pmatrix} \right\| = \left\| \begin{pmatrix} A^e_{11} \\ A^e_{21} \\ A^e_{31} \end{pmatrix} \right\| = \sqrt{(A^e_{11})^2 + (A^e_{21})^2 + (A^e_{31})^2} \\
2\, A^e_{\widehat{234}} &= \left\| \begin{pmatrix} \frac{\partial x}{\partial \eta} - \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \eta} - \frac{\partial y}{\partial \xi} \\ \frac{\partial z}{\partial \eta} - \frac{\partial z}{\partial \xi} \end{pmatrix} \times \begin{pmatrix} \frac{\partial x}{\partial \eta} - \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \eta} - \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \eta} - \frac{\partial z}{\partial \zeta} \end{pmatrix} \right\| = \left\| \begin{pmatrix} A^e_{10} \\ A^e_{20} \\ A^e_{30} \end{pmatrix} \right\| = \sqrt{(A^e_{10})^2 + (A^e_{20})^2 + (A^e_{30})^2}
\end{aligned} \tag{H.40}$$

Thus, the matrices and vectors of the ADE convective form (8.104) and (8.105) become for element e:

$$\begin{aligned}
\boldsymbol{O}^e &= \int_{\Omega^e} \acute{\mathcal{R}}^e \begin{pmatrix} N^e_1 N^e_1 & N^e_1 N^e_2 & N^e_1 N^e_3 & N^e_1 N^e_4 \\ N^e_2 N^e_1 & N^e_2 N^e_2 & N^e_2 N^e_3 & N^e_2 N^e_4 \\ N^e_3 N^e_1 & N^e_3 N^e_2 & N^e_3 N^e_3 & N^e_3 N^e_4 \\ N^e_4 N^e_1 & N^e_4 N^e_2 & N^e_4 N^e_3 & N^e_4 N^e_4 \end{pmatrix} d\Omega^e \\
&= \acute{\mathcal{R}}^e \int_0^1 \int_0^{1-\xi} \int_0^{1-\xi-\eta} \begin{pmatrix} (1-\xi-\eta-\zeta)^2 & (1-\xi-\eta-\zeta)\xi & (1-\xi-\eta-\zeta)\eta & (1-\xi-\eta-\zeta)\zeta \\ (1-\xi-\eta-\zeta)\xi & \xi^2 & \xi\eta & \xi\zeta \\ (1-\xi-\eta)\eta & \xi\eta & \eta^2 & \eta\zeta \\ (1-\xi-\eta)\zeta & \xi\zeta & \eta\zeta & \zeta^2 \end{pmatrix} \\
&\quad \times 6V^e d\zeta d\eta d\xi
\end{aligned}$$

$$= \frac{\acute{\mathcal{R}}^e V^e}{20} \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}$$

$$\boldsymbol{A}^e = \int_{\Omega^e} \begin{pmatrix} N_1^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_1^e(\boldsymbol{q}^e \cdot \nabla N_2^e) & N_1^e(\boldsymbol{q}^e \cdot \nabla N_3^e) & N_1^e(\boldsymbol{q}^e \cdot \nabla N_4^e) \\ N_2^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_2^e(\boldsymbol{q}^e \cdot \nabla N_2^e) & N_2^e(\boldsymbol{q}^e \cdot \nabla N_3^e) & N_2^e(\boldsymbol{q}^e \cdot \nabla N_4^e) \\ N_3^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_3^e(\boldsymbol{q}^e \cdot \nabla N_2^e) & N_3^e(\boldsymbol{q}^e \cdot \nabla N_3^e) & N_3^e(\boldsymbol{q}^e \cdot \nabla N_4^e) \\ N_4^e(\boldsymbol{q}^e \cdot \nabla N_1^e) & N_4^e(\boldsymbol{q}^e \cdot \nabla N_2^e) & N_4^e(\boldsymbol{q}^e \cdot \nabla N_3^e) & N_4^e(\boldsymbol{q}^e \cdot \nabla N_4^e) \end{pmatrix} d\Omega^e$$

$$= \frac{q^e}{6V^e} \int_0^1 \int_0^{1-\xi} \int_0^{1-\xi-\eta}$$

$$\times \begin{pmatrix} N_1^e(S_{11}^e + S_{21}^e + S_{31}^e) & N_1^e(S_{12}^e + S_{22}^e + S_{32}^e) & N_1^e(S_{13}^e + S_{23}^e + S_{33}^e) & N_1^e(S_{14}^e + S_{24}^e + S_{34}^e) \\ \xi(S_{11}^e + S_{21}^e + S_{31}^e) & \xi(S_{12}^e + S_{22}^e + S_{32}^e) & \xi(S_{13}^e + S_{23}^e + S_{33}^e) & \xi(S_{14}^e + S_{24}^e + S_{34}^e) \\ \eta(S_{11}^e + S_{21}^e + S_{31}^e) & \eta(S_{12}^e + S_{22}^e + S_{32}^e) & \eta(S_{13}^e + S_{23}^e + S_{33}^e) & \eta(S_{14}^e + S_{24}^e + S_{34}^e) \\ \zeta(S_{11}^e + S_{21}^e + S_{31}^e) & \zeta(S_{12}^e + S_{22}^e + S_{32}^e) & \zeta(S_{13}^e + S_{23}^e + S_{33}^e) & \zeta(S_{14}^e + S_{24}^e + S_{34}^e) \end{pmatrix}$$

$$\times 6V^e d\zeta d\eta d\xi$$

$$= \frac{q^e}{24} \begin{pmatrix} (S_{11}^e + S_{21}^e + S_{31}^e) & (S_{12}^e + S_{22}^e + S_{32}^e) & (S_{13}^e + S_{23}^e + S_{33}^e) & (S_{14}^e + S_{24}^e + S_{34}^e) \\ (S_{11}^e + S_{21}^e + S_{31}^e) & (S_{12}^e + S_{22}^e + S_{32}^e) & (S_{13}^e + S_{23}^e + S_{33}^e) & (S_{14}^e + S_{24}^e + S_{34}^e) \\ (S_{11}^e + S_{21}^e + S_{31}^e) & (S_{12}^e + S_{22}^e + S_{32}^e) & (S_{13}^e + S_{23}^e + S_{33}^e) & (S_{14}^e + S_{24}^e + S_{34}^e) \\ (S_{11}^e + S_{21}^e + S_{31}^e) & (S_{12}^e + S_{22}^e + S_{32}^e) & (S_{13}^e + S_{23}^e + S_{33}^e) & (S_{14}^e + S_{24}^e + S_{34}^e) \end{pmatrix}$$

$$\boldsymbol{C}^e = \int_{\Omega^e} \begin{pmatrix} \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) & \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_3^e) & \nabla N_1^e \cdot (\boldsymbol{D}^e \cdot \nabla N_4^e) \\ \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) & \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_3^e) & \nabla N_2^e \cdot (\boldsymbol{D}^e \cdot \nabla N_4^e) \\ \nabla N_3^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_3^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) & \nabla N_3^e \cdot (\boldsymbol{D}^e \cdot \nabla N_3^e) & \nabla N_3^e \cdot (\boldsymbol{D}^e \cdot \nabla N_4^e) \\ \nabla N_4^e \cdot (\boldsymbol{D}^e \cdot \nabla N_1^e) & \nabla N_4^e \cdot (\boldsymbol{D}^e \cdot \nabla N_2^e) & \nabla N_4^e \cdot (\boldsymbol{D}^e \cdot \nabla N_3^e) & \nabla N_4^e \cdot (\boldsymbol{D}^e \cdot \nabla N_4^e) \end{pmatrix}$$

$$\times d\Omega^e$$

$$= \frac{D^e}{(6V^e)^2} \int_0^1 \int_0^{1-\xi} \int_0^{1-\xi-\eta}$$

$$\times \begin{pmatrix} ((S_{11}^e)^2 + (S_{21}^e)^2 + (S_{31}^e)^2) & (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e + S_{31}^e S_{32}^e) & (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e + S_{31}^e S_{33}^e) & (S_{11}^e S_{14}^e + S_{21}^e S_{24}^e + S_{31}^e S_{34}^e) \\ (S_{12}^e S_{11}^e + S_{22}^e S_{21}^e + S_{32}^e S_{31}^e) & ((S_{12}^e)^2 + (S_{22}^e)^2 + (S_{32}^e)^2) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e + S_{32}^e S_{33}^e) & (S_{12}^e S_{14}^e + S_{22}^e S_{24}^e + S_{32}^e S_{34}^e) \\ (S_{13}^e S_{11}^e + S_{23}^e S_{21}^e + S_{33}^e S_{31}^e) & (S_{13}^e S_{12}^e + S_{23}^e S_{22}^e + S_{33}^e S_{32}^e) & ((S_{13}^e)^2 + (S_{23}^e)^2 + (S_{33}^e)^2) & (S_{13}^e S_{14}^e + S_{23}^e S_{24}^e + S_{33}^e S_{34}^e) \\ (S_{14}^e S_{11}^e + S_{24}^e S_{21}^e + S_{34}^e S_{31}^e) & (S_{14}^e S_{12}^e + S_{24}^e S_{22}^e + S_{34}^e S_{32}^e) & (S_{14}^e S_{13}^e + S_{24}^e S_{23}^e + S_{34}^e S_{33}^e) & ((S_{14}^e)^2 + (S_{24}^e)^2 + (S_{34}^e)^2) \end{pmatrix}$$

$$\times 6V^e d\zeta d\eta d\xi$$

$$= \frac{D^e}{36V^e}$$

$$\times \begin{pmatrix} ((S_{11}^e)^2 + (S_{21}^e)^2 + (S_{31}^e)^2) & (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e + S_{31}^e S_{32}^e) & (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e + S_{31}^e S_{33}^e) & (S_{11}^e S_{14}^e + S_{21}^e S_{24}^e + S_{31}^e S_{34}^e) \\ (S_{11}^e S_{12}^e + S_{21}^e S_{22}^e + S_{31}^e S_{32}^e) & ((S_{12}^e)^2 + (S_{22}^e)^2 + (S_{32}^e)^2) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e + S_{32}^e S_{33}^e) & (S_{12}^e S_{14}^e + S_{22}^e S_{24}^e + S_{32}^e S_{34}^e) \\ (S_{11}^e S_{13}^e + S_{21}^e S_{23}^e + S_{31}^e S_{33}^e) & (S_{12}^e S_{13}^e + S_{22}^e S_{23}^e + S_{32}^e S_{33}^e) & ((S_{13}^e)^2 + (S_{23}^e)^2 + (S_{33}^e)^2) & (S_{13}^e S_{14}^e + S_{23}^e S_{24}^e + S_{33}^e S_{34}^e) \\ (S_{11}^e S_{14}^e + S_{21}^e S_{24}^e + S_{31}^e S_{34}^e) & (S_{12}^e S_{14}^e + S_{22}^e S_{24}^e + S_{32}^e S_{34}^e) & (S_{13}^e S_{14}^e + S_{23}^e S_{24}^e + S_{33}^e S_{34}^e) & ((S_{14}^e)^2 + (S_{24}^e)^2 + (S_{34}^e)^2) \end{pmatrix}$$

$$\boldsymbol{R}^e = \int_{\Omega^e} (\vartheta^e + Q^e) \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e & N_1^e N_3^e & N_1^e N_4^e \\ N_2^e N_1^e & N_2^e N_2^e & N_2^e N_3^e & N_2^e N_4^e \\ N_3^e N_1^e & N_3^e N_2^e & N_3^e N_3^e & N_3^e N_4^e \\ N_4^e N_1^e & N_4^e N_2^e & N_4^e N_3^e & N_4^e N_4^e \end{pmatrix} d\Omega^e$$

$$= (\vartheta^e + Q^e) \int_0^1 \int_0^{1-\xi} \int_0^{1-\xi-\eta}$$

$$\times \begin{pmatrix} (1-\xi-\eta-\zeta)^2 & (1-\xi-\eta-\zeta)\xi & (1-\xi-\eta-\zeta)\eta & (1-\xi-\eta-\zeta)\zeta \\ (1-\xi-\eta-\zeta)\xi & \xi^2 & \xi\eta & \xi\zeta \\ (1-\xi-\eta)\eta & \xi\eta & \eta^2 & \eta\zeta \\ (1-\xi-\eta)\zeta & \xi\zeta & \eta\zeta & \zeta^2 \end{pmatrix} 6V^e d\zeta d\eta d\xi$$

$$= \frac{(\vartheta^e + Q^e) V^e}{20} \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix} \tag{H.41}$$

$$\boldsymbol{B}^e = \int_{\Gamma_C^e} \Phi^e \begin{pmatrix} N_1^e N_1^e & N_1^e N_2^e & N_1^e N_3^e & N_1^e N_4^e \\ N_2^e N_1^e & N_2^e N_2^e & N_2^e N_3^e & N_2^e N_4^e \\ N_3^e N_1^e & N_3^e N_2^e & N_3^e N_3^e & N_3^e N_4^e \\ N_4^e N_1^e & N_4^e N_2^e & N_4^e N_3^e & N_4^e N_4^e \end{pmatrix} d\Gamma^e$$

$$- \int_{\Gamma_{N_O}^e} \begin{pmatrix} N_1^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_1^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} & N_1^e(\boldsymbol{D} \cdot \nabla N_3^e) \cdot \boldsymbol{n} & N_1^e(\boldsymbol{D} \cdot \nabla N_4^e) \cdot \boldsymbol{n} \\ N_2^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_2^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} & N_2^e(\boldsymbol{D} \cdot \nabla N_3^e) \cdot \boldsymbol{n} & N_2^e(\boldsymbol{D} \cdot \nabla N_4^e) \cdot \boldsymbol{n} \\ N_3^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_3^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} & N_3^e(\boldsymbol{D} \cdot \nabla N_3^e) \cdot \boldsymbol{n} & N_3^e(\boldsymbol{D} \cdot \nabla N_4^e) \cdot \boldsymbol{n} \\ N_4^e(\boldsymbol{D} \cdot \nabla N_1^e) \cdot \boldsymbol{n} & N_4^e(\boldsymbol{D} \cdot \nabla N_2^e) \cdot \boldsymbol{n} & N_4^e(\boldsymbol{D} \cdot \nabla N_3^e) \cdot \boldsymbol{n} & N_4^e(\boldsymbol{D} \cdot \nabla N_4^e) \cdot \boldsymbol{n} \end{pmatrix}$$

$$\times d\Gamma^e$$

$$= \Phi^e \int_0^1 \int_0^{1-(\xi,\eta)}$$

$$\times \begin{pmatrix} (1-\xi-\eta-\zeta)^2 & (1-\xi-\eta-\zeta)\xi & (1-\xi-\eta-\zeta)\eta & (1-\xi-\eta-\zeta)\zeta \\ (1-\xi-\eta-\zeta)\xi & \xi^2 & \xi\eta & \xi\zeta \\ (1-\xi-\eta)\eta & \xi\eta & \eta^2 & \eta\zeta \\ (1-\xi-\eta)\zeta & \xi\zeta & \eta\zeta & \zeta^2 \end{pmatrix} \sqrt{\ldots}\, (d\eta d\xi, d\zeta d\xi, d\zeta d\eta)$$

$$- \frac{D^e}{6V^e} \int_0^1 \int_0^{1-(\xi,\eta)}$$

$$\times \begin{pmatrix} (1-\xi-\eta-\zeta) \sum_i S_{i1}^e n_i & (1-\xi-\eta-\zeta) \sum_i S_{i2}^e n_i & (1-\xi-\eta-\zeta) \sum_i S_{i3}^e n_i & (1-\xi-\eta-\zeta) \sum_i S_{i4}^e n_i \\ \xi \sum_i S_{i1}^e n_i & \xi \sum_i S_{i2}^e n_i & \xi \sum_i S_{i3}^e n_i & \xi \sum_i S_{i4}^e n_i \\ \eta \sum_i S_{i1}^e n_i & \eta \sum_i S_{i2}^e n_i & \eta \sum_i S_{i3}^e n_i & \eta \sum_i S_{i4}^e n_i \\ \zeta \sum_i S_{i1}^e n_i & \zeta \sum_i S_{i2}^e n_i & \zeta \sum_i S_{i3}^e n_i & \zeta \sum_i S_{i4}^e n_i \end{pmatrix}$$

$$\times \sqrt{\ldots}\, (d\eta d\xi, d\zeta d\xi, d\zeta d\eta)$$

$$
= \begin{cases}
\frac{\Phi^e}{12}\begin{pmatrix} 2&1&1&0\\ 1&2&1&0\\ 1&1&2&0\\ 0&0&0&0 \end{pmatrix} A^e_{\widehat{123}} - \\
\frac{D^e}{18V^e}\begin{pmatrix} \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ 0&0&0&0 \end{pmatrix}_{\Gamma_{N_O}} A^e_{\widehat{123}} + & \text{on area } \widehat{123} \\
\frac{\Phi^e}{12}\begin{pmatrix} 2&1&0&1\\ 1&2&0&1\\ 0&0&0&0\\ 1&1&0&2 \end{pmatrix} A^e_{\widehat{124}} - \\
\frac{D^e}{18V^e}\begin{pmatrix} \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ 0&0&0&0 \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \end{pmatrix}_{\Gamma_{N_O}} A^e_{\widehat{124}} + & \text{on area } \widehat{124} \\
\frac{\Phi^e}{12}\begin{pmatrix} 2&0&1&1\\ 0&0&0&0\\ 1&0&2&1\\ 1&0&1&2 \end{pmatrix} A^e_{\widehat{134}} - \\
\frac{D^e}{18V^e}\begin{pmatrix} \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ 0&0&0&0 \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \end{pmatrix}_{\Gamma_{N_O}} A^e_{\widehat{134}} + & \text{on edge } \widehat{134} \\
\frac{\Phi^e}{12}\begin{pmatrix} 0&0&0&0\\ 0&2&1&1\\ 0&1&2&1\\ 0&1&1&2 \end{pmatrix} A^e_{\widehat{234}} - \\
\frac{D^e}{18V^e}\begin{pmatrix} 0&0&0&0 \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \end{pmatrix}_{\Gamma_{N_O}} A^e_{\widehat{234}} & \text{on edge } \widehat{234}
\end{cases}
\tag{H.42}
$$

with $\sum_i S^e_{i1}n_i = S^e_{11}n_1 + S^e_{21}n_2 + S^e_{31}n_3$ and so forth,

$$
\begin{aligned}
\boldsymbol{H}^e &= \int_{\Gamma^e_C} \Phi^e \phi^e_C \begin{pmatrix} N^e_1 \\ N^e_2 \\ N^e_3 \\ N^e_4 \end{pmatrix} d\Gamma^e - \int_{\Gamma^e_N} q^e_N \begin{pmatrix} N^e_1 \\ N^e_2 \\ N^e_3 \\ N^e_4 \end{pmatrix} d\Gamma^e \\
&= \Phi^e \phi^e_C \int_0^1 \int_0^{1-(\xi,\eta)} \begin{pmatrix} 1-\xi-\eta-\zeta \\ \xi \\ \eta \\ \zeta \end{pmatrix} \sqrt{\ldots}\, (d\eta d\xi, d\zeta d\xi, d\zeta d\eta) \\
&\quad -q^e_N \int_0^1 \int_0^{1-(\xi,\eta)} \begin{pmatrix} 1-\xi-\eta-\zeta \\ \xi \\ \eta \\ \zeta \end{pmatrix} \sqrt{\ldots}\, (d\eta d\xi, d\zeta d\xi, d\zeta d\eta)
\end{aligned}
$$

$$= \begin{cases} \left[\frac{\Phi^e\phi_C^e}{3}\begin{pmatrix}1\\1\\1\\0\end{pmatrix} - \frac{q_N^e}{3}\begin{pmatrix}1\\1\\1\\0\end{pmatrix}\right]A^e_{\widehat{123}}+ & \text{on area } \widehat{123} \\ \left[\frac{\Phi^e\phi_C^e}{3}\begin{pmatrix}1\\1\\0\\1\end{pmatrix} - \frac{q_N^e}{3}\begin{pmatrix}1\\1\\0\\1\end{pmatrix}\right]A^e_{\widehat{124}}+ & \text{on area } \widehat{124} \\ \left[\frac{\Phi^e\phi_C^e}{3}\begin{pmatrix}1\\0\\1\\1\end{pmatrix} - \frac{q_N^e}{3}\begin{pmatrix}1\\0\\1\\1\end{pmatrix}\right]A^e_{\widehat{134}}+ & \text{on area } \widehat{134} \\ \left[\frac{\Phi^e\phi_C^e}{3}\begin{pmatrix}0\\1\\1\\1\end{pmatrix} - \frac{q_N^e}{3}\begin{pmatrix}0\\1\\1\\1\end{pmatrix}\right]A^e_{\widehat{234}} & \text{on area } \widehat{234} \end{cases} \tag{H.43}$$

$$\begin{aligned} \boldsymbol{Q}^e &= \int_{\Omega^e} H^e \begin{pmatrix}N_1^e\\N_2^e\\N_3^e\\N_4^e\end{pmatrix} d\Omega^e - \textstyle\sum_w \phi_w(\boldsymbol{x}_w)Q_w(t) \\ &= H^e \int_0^1 \int_0^{1-\xi} \int_0^{1-\xi-\eta} \begin{pmatrix}1-\xi-\eta-\zeta\\ \xi\\ \eta\\ \zeta\end{pmatrix} 6V^e d\zeta d\eta d\xi - \textstyle\sum_w \phi_w(\boldsymbol{x}_w)Q_w(t) \\ &= \frac{H^eV^e}{4}\begin{pmatrix}1\\1\\1\\1\end{pmatrix} - \textstyle\sum_w \phi_w(\boldsymbol{x}_w)Q_w(t) \end{aligned} \tag{H.44}$$

where for convenience we assume constant parameters (storage coefficient $\acute{\mathcal{R}}^e$, flux $\boldsymbol{q}^e$, dispersion $\boldsymbol{D}^e$, decay rate ϑ^e, flow supply Q^e, transfer coefficient Φ^e and source/sink H^e) within the element. In (H.42) the components of the unit normal vector $\boldsymbol{n}$ in the Cartesian $x-$, $y-$ and $z-$coordinate direction on the outflow boundary Γ_{N_O} are indicated by $n1$, n_2 and n_3, respectively. Finally, the spatially discretized ADE in the convective form (8.99) can be summarized as

$$\sum_e \Bigg\{ \frac{\acute{\mathcal{R}}^eV^e}{20}\begin{pmatrix}2&1&1&1\\1&2&1&1\\1&1&2&1\\1&1&1&2\end{pmatrix} \cdot \begin{pmatrix}\frac{d\phi_1^e}{dt}\\ \frac{d\phi_2^e}{dt}\\ \frac{d\phi_3^e}{dt}\\ \frac{d\phi_4^e}{dt}\end{pmatrix}$$

$$+ \Bigg[\frac{q^e}{24}\begin{pmatrix} (S_{11}^e+S_{21}^e+S_{31}^e) & (S_{12}^e+S_{22}^e+S_{32}^e) & (S_{13}^e+S_{23}^e+S_{33}^e) & (S_{14}^e+S_{24}^e+S_{34}^e) \\ (S_{11}^e+S_{21}^e+S_{31}^e) & (S_{12}^e+S_{22}^e+S_{32}^e) & (S_{13}^e+S_{23}^e+S_{33}^e) & (S_{14}^e+S_{24}^e+S_{34}^e) \\ (S_{11}^e+S_{21}^e+S_{31}^e) & (S_{12}^e+S_{22}^e+S_{32}^e) & (S_{13}^e+S_{23}^e+S_{33}^e) & (S_{14}^e+S_{24}^e+S_{34}^e) \\ (S_{11}^e+S_{21}^e+S_{31}^e) & (S_{12}^e+S_{22}^e+S_{32}^e) & (S_{13}^e+S_{23}^e+S_{33}^e) & (S_{14}^e+S_{24}^e+S_{34}^e) \end{pmatrix}$$

$$+ \tfrac{D^e}{36V^e}$$

$$\times \begin{pmatrix} ((S^e_{11})^2+(S^e_{21})^2+(S^e_{31})^2) & (S^e_{11}S^e_{12}+S^e_{21}S^e_{22}+S^e_{31}S^e_{32}) & (S^e_{11}S^e_{13}+S^e_{21}S^e_{23}+S^e_{31}S^e_{33}) & (S^e_{11}S^e_{14}+S^e_{21}S^e_{24}+S^e_{31}S^e_{34}) \\ (S^e_{11}S^e_{12}+S^e_{21}S^e_{22}+S^e_{31}S^e_{32}) & ((S^e_{12})^2+(S^e_{22})^2+(S^e_{32})^2) & (S^e_{12}S^e_{13}+S^e_{22}S^e_{23}+S^e_{32}S^e_{33}) & (S^e_{12}S^e_{14}+S^e_{22}S^e_{24}+S^e_{32}S^e_{34}) \\ (S^e_{11}S^e_{13}+S^e_{21}S^e_{23}+S^e_{31}S^e_{33}) & (S^e_{12}S^e_{13}+S^e_{22}S^e_{23}+S^e_{32}S^e_{33}) & ((S^e_{13})^2+(S^e_{23})^2+(S^e_{33})^2) & (S^e_{13}S^e_{14}+S^e_{23}S^e_{24}+S^e_{33}S^e_{34}) \\ (S^e_{11}S^e_{14}+S^e_{21}S^e_{24}+S^e_{31}S^e_{34}) & (S^e_{12}S^e_{14}+S^e_{22}S^e_{24}+S^e_{32}S^e_{34}) & (S^e_{13}S^e_{14}+S^e_{23}S^e_{24}+S^e_{33}S^e_{34}) & ((S^e_{14})^2+(S^e_{24})^2+(S^e_{34})^2) \end{pmatrix}$$

$$+ \tfrac{(\vartheta^e+Q^e)V^e}{20} \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix} + \frac{\Phi^e}{12} \Bigg(\begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} A^e_{\widehat{123}} + \begin{pmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix} A^e_{\widehat{124}} + \begin{pmatrix} 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix} A^e_{\widehat{134}}$$

$$+ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix} A^e_{\widehat{234}} \Bigg) - \frac{D^e}{18V^e} \Bigg(\begin{pmatrix} \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ 0 & 0 & 0 & 0 \end{pmatrix} A^e_{\widehat{123}}$$

$$+ \begin{pmatrix} \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ 0 & 0 & 0 & 0 \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \end{pmatrix} A^e_{\widehat{124}}$$

$$+ \begin{pmatrix} \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ 0 & 0 & 0 & 0 \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \end{pmatrix} A^e_{\widehat{134}}$$

$$+ \begin{pmatrix} 0 & 0 & 0 & 0 \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \\ \sum_i S^e_{i1}n_i & \sum_i S^e_{i2}n_i & \sum_i S^e_{i3}n_i & \sum_i S^e_{i4}n_i \end{pmatrix} A^e_{\widehat{234}} \Bigg)_{\Gamma_{N_O}} \Bigg] \cdot \begin{pmatrix} \phi^e_1 \\ \phi^e_2 \\ \phi^e_3 \\ \phi^e_4 \end{pmatrix}$$

$$- \frac{\Phi^e \phi^e_C - q^e_N}{3} \left[\begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} A^e_{\widehat{123}} + \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} A^e_{\widehat{124}} + \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} A^e_{\widehat{134}} + \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} A^e_{\widehat{234}} \right]$$

$$- \frac{H^e V^e}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \Bigg\} + \sum_w \phi_w(\boldsymbol{x}_w) Q_w(t) = \mathbf{0} \tag{H.45}$$

Similar expressions can be obtained for the divergence form of ADE (8.98).

H.4 Simplified Element Shapes in 2D and 3D to Attain Constant Jacobians

H.4.1 Linear Quadrilateral Element as Parallelogram or Rectangle

The linear 4-node quadrilateral element as described in Appendix G (Table G.2b) has the following shape functions

$$\begin{aligned} N_1^e &= \tfrac{1}{4}(1-\xi)(1-\eta) \\ N_2^e &= \tfrac{1}{4}(1+\xi)(1-\eta) \\ N_3^e &= \tfrac{1}{4}(1+\xi)(1+\eta) \\ N_4^e &= \tfrac{1}{4}(1-\xi)(1+\eta) \end{aligned} \tag{H.46}$$

and the local derivatives

$$\begin{aligned} \frac{\partial N_1^e}{\partial \xi} &= -\tfrac{1}{4}(1-\eta), & \frac{\partial N_1^e}{\partial \eta} &= -\tfrac{1}{4}(1-\xi) \\ \frac{\partial N_2^e}{\partial \xi} &= \tfrac{1}{4}(1-\eta), & \frac{\partial N_2^e}{\partial \eta} &= -\tfrac{1}{4}(1+\xi) \\ \frac{\partial N_3^e}{\partial \xi} &= \tfrac{1}{4}(1+\eta), & \frac{\partial N_3^e}{\partial \eta} &= \tfrac{1}{4}(1+\xi) \\ \frac{\partial N_4^e}{\partial \xi} &= -\tfrac{1}{4}(1+\eta), & \frac{\partial N_4^e}{\partial \eta} &= \tfrac{1}{4}(1-\xi) \end{aligned} \tag{H.47}$$

for $(-1 \leq \xi, \eta \leq 1)$. It is for the quadrilateral element, cf. (8.71)

$$\begin{aligned} x &= N_1^e x_1^e + N_2^e x_2^e + N_3^e x_3^e + N_4^e x_4^e \\ y &= N_1^e y_1^e + N_2^e y_2^e + N_3^e y_3^e + N_4^e y_4^e \end{aligned} \tag{H.48}$$

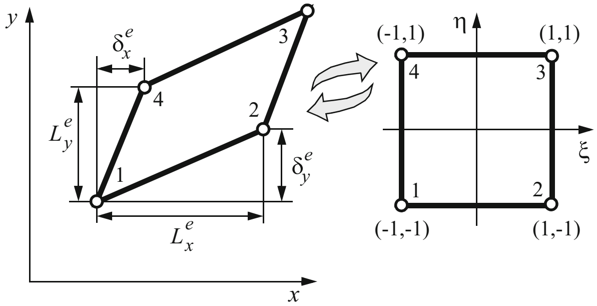
where x_I^e, y_I^e $(I = 1, 2, 3, 4)$ correspond to the Cartesian coordinates of the vertices (nodes) of the quadrilateral. Using (H.47) and (H.48) in (8.116) we obtain the Jacobian

$$\boldsymbol{J}^e = \frac{1}{4}\begin{pmatrix} (-x_1^e + x_2^e + x_3^e - x_4^e + \eta(x_1^e - x_2^e + x_3^e - x_4^e)) & (-y_1^e + y_2^e + y_3^e - y_4^e + \eta(y_1^e - y_2^e + y_3^e - y_4^e)) \\ (-x_1^e - x_2^e + x_3^e + x_4^e + \xi(x_1^e - x_2^e + x_3^e - x_4^e)) & (-y_1^e - y_2^e + y_3^e + y_4^e + \xi(y_1^e - y_2^e + y_3^e - y_4^e)) \end{pmatrix} \tag{H.49}$$

We see from (H.49) that the Jacobian $\boldsymbol{J}^e$ of the quadrilateral is a function of the local coordinates ξ, η, which makes an analytical evaluation impossible. To eliminate the ξ, η–dependency from $\boldsymbol{J}^e$ we have to enforce $x_1^e - x_2^e + x_3^e - x_4^e = 0$ and $y_1^e - y_2^e + y_3^e - y_4^e = 0$, which become valid for the parallelogram and the rectangle as shown in Fig. H.4. The parallelogram is characterized by the geometric conditions:

$$\begin{aligned} \delta_x^e &= x_4^e - x_1^e = x_3^e - x_2^e & L_x^e &= x_2^e - x_1^e = x_3^e - x_4^e \\ \delta_y^e &= y_2^e - y_1^e = y_3^e - y_4^e & L_y^e &= y_4^e - y_1^e = y_3^e - y_2^e \end{aligned} \tag{H.50}$$

Fig. H.4 Linear 4-node parallelogram in the global and local coordinate system



Then, the Jacobain of the parallelogram

$$\boldsymbol{J}^e = \tfrac{1}{2}\begin{pmatrix} L_x^e & \delta_y^e \\ \delta_x^e & L_y^e \end{pmatrix} \tag{H.51}$$

becomes constant. It yields the inverse Jacobian

$$(\boldsymbol{J}^e)^{-1} = \frac{1}{2|\boldsymbol{J}^e|}\begin{pmatrix} L_y^e & -\delta_y^e \\ -\delta_x^e & L_x^e \end{pmatrix} \tag{H.52}$$

and the determinant

$$|\boldsymbol{J}^e| = \tfrac{1}{4}(L_x^e L_y^e - \delta_x^e \delta_y^e) = \tfrac{1}{4}A^e \tag{H.53}$$

which corresponds to the quarter of the parallelogram area. We note that $\delta_x^e = \delta_y^e = 0$ for a rectangular element aligned to the global coordinate axes. Now, it is possible to evaluate the element matrices and vectors for the parallelogram and the rectangle in a similar way as done in Sect. H.2 for the triangle.

H.4.2 Linear Hexahedral Element as Parallelepiped or Brick

The linear 8-node hexahedral element as described in Appendix G (Table G.3c) has the following shape functions

$$\begin{aligned}
N_1^e &= \tfrac{1}{8}(1-\xi)(1-\eta)(1+\zeta) \\
N_2^e &= \tfrac{1}{8}(1+\xi)(1-\eta)(1+\zeta) \\
N_3^e &= \tfrac{1}{8}(1+\xi)(1+\eta)(1+\zeta) \\
N_4^e &= \tfrac{1}{8}(1-\xi)(1+\eta)(1+\zeta)
\end{aligned}$$

$$\begin{aligned}
N_5^e &= \tfrac{1}{8}(1-\xi)(1-\eta)(1-\zeta)\\
N_6^e &= \tfrac{1}{8}(1+\xi)(1-\eta)(1-\zeta)\\
N_7^e &= \tfrac{1}{8}(1+\xi)(1+\eta)(1-\zeta)\\
N_8^e &= \tfrac{1}{8}(1-\xi)(1+\eta)(1-\zeta)
\end{aligned} \tag{H.54}$$

and the local derivatives

$$\begin{array}{lll}
\frac{\partial N_1^e}{\partial \xi} = -\frac{1}{8}(1-\eta)(1+\zeta), & \frac{\partial N_1^e}{\partial \eta} = -\frac{1}{8}(1-\xi)(1+\zeta), & \frac{\partial N_1^e}{\partial \zeta} = \frac{1}{8}(1-\xi)(1-\eta)\\
\frac{\partial N_2^e}{\partial \xi} = \frac{1}{8}(1-\eta)(1+\zeta), & \frac{\partial N_2^e}{\partial \eta} = -\frac{1}{8}(1+\xi)(1+\zeta), & \frac{\partial N_2^e}{\partial \zeta} = \frac{1}{8}(1+\xi)(1-\eta)\\
\frac{\partial N_3^e}{\partial \xi} = \frac{1}{8}(1+\eta)(1+\zeta), & \frac{\partial N_3^e}{\partial \eta} = \frac{1}{8}(1+\xi)(1+\zeta), & \frac{\partial N_3^e}{\partial \zeta} = \frac{1}{8}(1+\xi)(1+\eta)\\
\frac{\partial N_4^e}{\partial \xi} = -\frac{1}{8}(1+\eta)(1+\zeta), & \frac{\partial N_4^e}{\partial \eta} = \frac{1}{8}(1-\xi)(1+\zeta), & \frac{\partial N_4^e}{\partial \zeta} = \frac{1}{8}(1-\xi)(1+\eta)\\
\frac{\partial N_5^e}{\partial \xi} = -\frac{1}{8}(1-\eta)(1-\zeta), & \frac{\partial N_5^e}{\partial \eta} = -\frac{1}{8}(1-\xi)(1-\zeta), & \frac{\partial N_5^e}{\partial \zeta} = -\frac{1}{8}(1-\xi)(1-\eta)\\
\frac{\partial N_6^e}{\partial \xi} = \frac{1}{8}(1-\eta)(1-\zeta), & \frac{\partial N_6^e}{\partial \eta} = -\frac{1}{8}(1+\xi)(1-\zeta), & \frac{\partial N_6^e}{\partial \zeta} = -\frac{1}{8}(1+\xi)(1-\eta)\\
\frac{\partial N_7^e}{\partial \xi} = \frac{1}{8}(1+\eta)(1-\zeta), & \frac{\partial N_7^e}{\partial \eta} = \frac{1}{8}(1+\xi)(1-\zeta), & \frac{\partial N_7^e}{\partial \zeta} = -\frac{1}{8}(1+\xi)(1+\eta)\\
\frac{\partial N_8^e}{\partial \xi} = -\frac{1}{8}(1+\eta)(1-\zeta), & \frac{\partial N_8^e}{\partial \eta} = \frac{1}{8}(1-\xi)(1-\zeta), & \frac{\partial N_8^e}{\partial \zeta} = -\frac{1}{8}(1-\xi)(1+\eta)
\end{array} \tag{H.55}$$

for $(-1 \leq \xi, \eta, \zeta \leq 1)$. It is for the hexahedral element, cf. (8.71)

$$x = \sum_{I=1}^{8} N_I^e x_I^e, \quad y = \sum_{I=1}^{8} N_I^e y_I^e, \quad z = \sum_{I=1}^{8} N_I^e z_I^e \tag{H.56}$$

where x_I^e, y_I^e, z_I^e $(I = 1, \ldots, 8)$ correspond to the Cartesian coordinates of the vertices (nodes) of the hexahedron. Using (H.55) and (H.56) in (8.115) we obtain the Jacobian

$$\boldsymbol{J}^e = \frac{1}{8}\begin{pmatrix}
(a_{\xi x} + \eta b_x + \zeta c_x + \eta\zeta d_x) & (a_{\xi y} + \eta b_y + \zeta c_y + \eta\zeta d_y) & (a_{\xi z} + \eta b_z + \zeta c_z + \eta\zeta d_z)\\
(a_{\eta x} + \xi b_x + \zeta e_x + \xi\zeta d_x) & (a_{\eta y} + \xi b_y + \zeta e_y + \xi\zeta d_y) & (a_{\eta z} + \xi b_z + \zeta e_z + \xi\zeta d_z)\\
(a_{\zeta x} + \xi c_x + \eta e_x + \xi\eta b_x) & (a_{\zeta y} + \xi c_y + \eta e_y + \xi\eta b_y) & (a_{\zeta z} + \xi c_z + \eta e_z + \xi\eta b_z)
\end{pmatrix} \tag{H.57}$$

where

$$\left.\begin{aligned}
a_{\xi x} &= -x_1^e + x_2^e + x_3^e - x_4^e - x_5^e + x_6^e + x_7^e - x_8^e\\
a_{\xi y} &= -y_1^e + y_2^e + y_3^e - y_4^e - y_5^e + y_6^e + y_7^e - y_8^e\\
a_{\xi z} &= -z_1^e + z_2^e + z_3^e - z_4^e - z_5^e + z_6^e + z_7^e - z_8^e\\
a_{\eta x} &= -x_1^e - x_2^e + x_3^e + x_4^e - x_5^e - x_6^e + x_7^e + x_8^e\\
a_{\eta y} &= -y_1^e - y_2^e + y_3^e + y_4^e - y_5^e - y_6^e + y_7^e + y_8^e\\
a_{\eta z} &= -z_1^e - z_2^e + z_3^e + z_4^e - z_5^e - z_6^e + z_7^e + z_8^e\\
a_{\zeta x} &= x_1^e + x_2^e + x_3^e + x_4^e - x_5^e - x_6^e - x_7^e - x_8^e\\
a_{\zeta y} &= y_1^e + y_2^e + y_3^e + y_4^e - y_5^e - y_6^e - y_7^e - y_8^e\\
a_{\zeta z} &= z_1^e + z_2^e + z_3^e + z_4^e - z_5^e - z_6^e - z_7^e - z_8^e
\end{aligned}\right\} \tag{H.58}$$

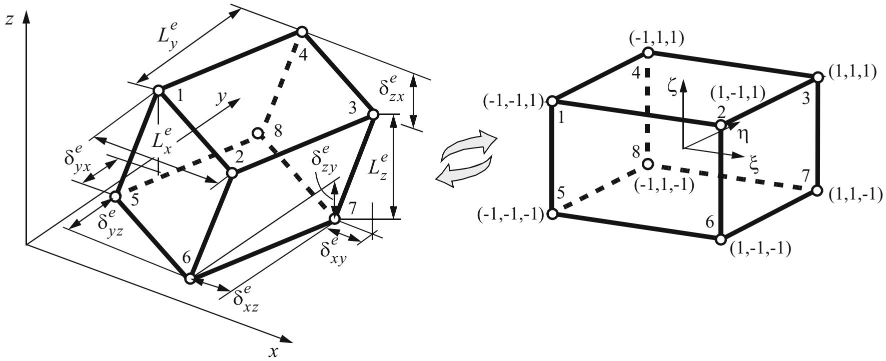


Fig. H.5 Linear 8-node parallelepiped in the global and local coordinate system

$$\left.\begin{aligned}
b_x &= x_1^e - x_2^e + x_3^e - x_4^e + x_5^e - x_6^e + x_7^e - x_8^e \\
b_y &= y_1^e - y_2^e + y_3^e - y_4^e + y_5^e - y_6^e + y_7^e - y_8^e \\
b_z &= z_1^e - z_2^e + z_3^e - z_4^e + z_5^e - z_6^e + z_7^e - z_8^e \\
c_x &= -x_1^e + x_2^e + x_3^e - x_4^e + x_5^e - x_6^e - x_7^e + x_8^e \\
c_y &= -y_1^e + y_2^e + y_3^e - y_4^e + y_5^e - y_6^e - y_7^e + y_8^e \\
c_z &= -z_1^e + z_2^e + z_3^e - z_4^e + z_5^e - z_6^e - z_7^e + z_8^e \\
d_x &= x_1^e - x_2^e + x_3^e - x_4^e - x_5^e + x_6^e - x_7^e + x_8^e \\
d_y &= y_1^e - y_2^e + y_3^e - y_4^e - y_5^e + y_6^e - y_7^e + y_8^e \\
d_z &= z_1^e - z_2^e + z_3^e - z_4^e - z_5^e + z_6^e - z_7^e + z_8^e \\
e_x &= -x_1^e - x_2^e + x_3^e + x_4^e + x_5^e + x_6^e - x_7^e - x_8^e \\
e_y &= -y_1^e - y_2^e + y_3^e + y_4^e + y_5^e + y_6^e - y_7^e - y_8^e \\
e_z &= -z_1^e - z_2^e + z_3^e + z_4^e + z_5^e + z_6^e - z_7^e - z_8^e
\end{aligned}\right\} \tag{H.59}$$

It is obvious from (H.57) that the Jacobian $\boldsymbol{J}^e$ of the hexahedron is still a function of the local coordinates ξ, η, ζ. Any analytical evaluation needs that the ξ, η, ζ—dependencies in $\boldsymbol{J}^e$ vanish, i.e., all coefficients $b,\ c,\ d$ and e of (H.59) must be zero. This is valid for the parallelepiped and the brick as shown in Fig. H.5, where the following geometric conditions hold:

$$\left.\begin{aligned}
\delta_{xz}^e &= x_4^e - x_1^e = x_3^e - x_2^e = x_8^e - x_5^e = x_7^e - x_6^e \\
\delta_{xy}^e &= x_4^e - x_8^e = x_1^e - x_5^e = x_3^e - x_7^e = x_2^e - x_6^e \\
\delta_{yx}^e &= y_1^e - y_5^e = y_2^e - y_6^e = y_4^e - y_8^e = y_3^e - y_7^e \\
\delta_{yz}^e &= y_2^e - y_1^e = y_3^e - y_4^e = y_6^e - y_5^e = y_7^e - y_8^e \\
\delta_{zx}^e &= z_2^e - z_1^e = z_6^e - z_5^e = z_3^e - z_4^e = z_7^e - z_8^e \\
\delta_{zy}^e &= z_8^e - z_5^e = z_4^e - z_1^e = z_3^e - z_2^e = z_7^e - z_6^e \\
L_x^e &= x_2^e - x_1^e = x_3^e - x_4^e = x_6^e - x_5^e = x_7^e - x_8^e \\
L_y^e &= y_4^e - y_1^e = y_3^e - y_2^e = y_8^e - y_5^e = y_7^e - y_6^e \\
L_z^e &= z_4^e - z_8^e = z_1^e - z_5^e = z_3^e - z_7^e = z_2^e - z_6^e
\end{aligned}\right\} \tag{H.60}$$

Then, the Jacobain of the parallelepiped

$$\boldsymbol{J}^e = \tfrac{1}{2}\begin{pmatrix} L_x^e & \delta_{yz}^e & \delta_{zx}^e \\ \delta_{xz}^e & L_y^e & \delta_{zy}^e \\ \delta_{xy}^e & \delta_{yx}^e & L_z^e \end{pmatrix} \tag{H.61}$$

becomes constant. It yields the inverse Jacobian

$$(\boldsymbol{J}^e)^{-1} = \frac{1}{4|\boldsymbol{J}^e|}\begin{pmatrix} (L_y^e L_z^e - \delta_{yx}^e\delta_{zy}^e) & (\delta_{zx}^e\delta_{yx}^e - \delta_{yz}^e L_z^e) & (\delta_{yz}^e\delta_{zy}^e - \delta_{zx}^e L_y^e) \\ (\delta_{xy}^e\delta_{zy}^e - \delta_{xz}^e L_z^e) & (L_x^e L_z^e - \delta_{zx}^e\delta_{xy}^e) & (\delta_{xz}^e\delta_{zx}^e - \delta_{zy}^e L_x^e) \\ (\delta_{xz}^e\delta_{yx}^e - \delta_{xy}^e L_y^e) & (\delta_{yz}^e\delta_{xy}^e - \delta_{yx}^e L_x^e) & (L_x^e L_y^e - \delta_{yz}^e\delta_{xz}^e) \end{pmatrix} \tag{H.62}$$

and the determinant

$$|\boldsymbol{J}^e| = \tfrac{1}{8}[L_x^e(L_y^e L_z^e - \delta_{yx}^e\delta_{zy}^e) + \delta_{xz}^e(\delta_{yx}^e\delta_{zx}^e - \delta_{yz}^e L_z^e) + \delta_{xy}^e(\delta_{yz}^e\delta_{zy}^e - \delta_{zx}^e L_y^e)] = \tfrac{1}{8}V^e \tag{H.63}$$

which corresponds to one-eighth of the parallelepiped volume. We note that $\delta_{xz}^e = \delta_{xy}^e = \delta_{yx}^e = \delta_{yz}^e = \delta_{zx}^e = \delta_{zy}^e = 0$ for a brick element aligned to the global coordinate axes. The element matrices and vectors can now be analytically evaluated for the parallelepiped and the brick element in a similar way as done in Sect. H.3 for the tetrahedron.

H.4.3 Linear Pentahedral Element as Triangular Prism with Parallel Top and Bottom Surfaces

The linear 6-node pentahedral element as described in Appendix G (Table G.3b) has the following shape functions

$$\begin{aligned} N_1^e &= \tfrac{1}{2}(1-\xi-\eta)(1+\zeta) \\ N_2^e &= \tfrac{1}{2}\xi(1+\zeta) \\ N_3^e &= \tfrac{1}{2}\eta(1+\zeta) \\ N_4^e &= \tfrac{1}{2}(1-\xi-\eta)(1-\zeta) \\ N_5^e &= \tfrac{1}{2}\xi(1-\zeta) \\ N_6^e &= \tfrac{1}{2}\eta(1-\zeta) \end{aligned} \tag{H.64}$$

and the local derivatives

$$\begin{aligned} &\tfrac{\partial N_1^e}{\partial \xi} = -\tfrac{1}{2}(1+\zeta), \quad &&\tfrac{\partial N_1^e}{\partial \eta} = -\tfrac{1}{2}(1+\zeta), \quad &&\tfrac{\partial N_1^e}{\partial \zeta} = \tfrac{1}{2}(1-\xi-\eta) \\ &\tfrac{\partial N_2^e}{\partial \xi} = \tfrac{1}{2}(1+\zeta), &&\tfrac{\partial N_2^e}{\partial \eta} = 0, &&\tfrac{\partial N_2^e}{\partial \zeta} = \tfrac{1}{2}\xi \end{aligned}$$

$$\begin{aligned}
&\frac{\partial N_3^e}{\partial \xi} = 0, && \frac{\partial N_3^e}{\partial \eta} = \tfrac{1}{2}(1+\zeta), && \frac{\partial N_3^e}{\partial \zeta} = \tfrac{1}{2}\eta \\
&\frac{\partial N_4^e}{\partial \xi} = -\tfrac{1}{2}(1-\zeta), && \frac{\partial N_4^e}{\partial \eta} = -\tfrac{1}{2}(1-\zeta), && \frac{\partial N_4^e}{\partial \zeta} = -\tfrac{1}{2}(1-\xi-\eta) \\
&\frac{\partial N_5^e}{\partial \xi} = \tfrac{1}{2}(1-\zeta), && \frac{\partial N_5^e}{\partial \eta} = 0, && \frac{\partial N_5^e}{\partial \zeta} = -\tfrac{1}{2}\xi \\
&\frac{\partial N_6^e}{\partial \xi} = 0, && \frac{\partial N_6^e}{\partial \eta} = \tfrac{1}{2}(1-\zeta), && \frac{\partial N_6^e}{\partial \zeta} = -\tfrac{1}{2}\eta
\end{aligned} \tag{H.65}$$

for $(0 \leq \xi, \eta \leq 1)$ and $(-1 \leq \zeta \leq 1)$. It is for the pentahedral element, cf. (8.71)

$$x = \sum_{I=1}^{6} N_I^e x_I^e, \quad y = \sum_{I=1}^{6} N_I^e y_I^e, \quad z = \sum_{I=1}^{6} N_I^e z_I^e \tag{H.66}$$

where x_I^e, y_I^e, z_I^e $(I = 1, \ldots, 6)$ correspond to the Cartesian coordinates of the vertices (nodes) of the pentahedron. Using (H.65) and (H.66) in (8.115) we obtain the Jacobian

$$\boldsymbol{J}^e = \frac{1}{2}\begin{pmatrix} (a_{\xi x} + \zeta b_x) & (a_{\xi y} + \zeta b_y) & (a_{\xi z} + \zeta b_z) \\ (a_{\eta x} + \zeta c_x) & (a_{\eta y} + \zeta c_y) & (a_{\eta z} + \zeta c_z) \\ (a_{\zeta x} + \xi b_x + \eta c_x) & (a_{\zeta y} + \xi b_y + \eta c_y) & (a_{\zeta z} + \xi b_z + \eta c_z) \end{pmatrix} \tag{H.67}$$

where

$$\left.\begin{aligned}
a_{\xi x} &= -x_1^e + x_2^e - x_4^e + x_5^e \\
a_{\xi y} &= -y_1^e + y_2^e - y_4^e + y_5^e \\
a_{\xi z} &= -z_1^e + z_2^e - z_4^e + z_5^e \\
a_{\eta x} &= -x_1^e + x_3^e - x_4^e + x_6^e \\
a_{\eta y} &= -y_1^e + y_3^e - y_4^e + y_6^e \\
a_{\eta z} &= -z_1^e + z_3^e - z_4^e + z_6^e \\
a_{\zeta x} &= x_1^e - x_4^e \\
a_{\zeta y} &= y_1^e - y_4^e \\
a_{\zeta z} &= z_1^e - z_4^e
\end{aligned}\right\} \tag{H.68}$$

$$\left.\begin{aligned}
b_x &= -x_1^e + x_2^e + x_4^e - x_5^e \\
b_y &= -y_1^e + y_2^e + y_4^e - y_5^e \\
b_z &= -z_1^e + z_2^e + z_4^e - z_5^e \\
c_x &= -x_1^e + x_3^e + x_4^e - x_6^e \\
c_y &= -y_1^e + y_3^e + y_4^e - y_6^e \\
c_z &= -z_1^e + z_3^e + z_4^e - z_6^e
\end{aligned}\right\} \tag{H.69}$$

It is obvious from (H.67) that the Jacobian $\boldsymbol{J}^e$ of the hexahedron is still a function of the local coordinates ξ, η, ζ. Any analytical evaluation needs that the ξ, η, ζ−dependencies in $\boldsymbol{J}^e$ vanish, i.e., all coefficients b and c of (H.69) must be

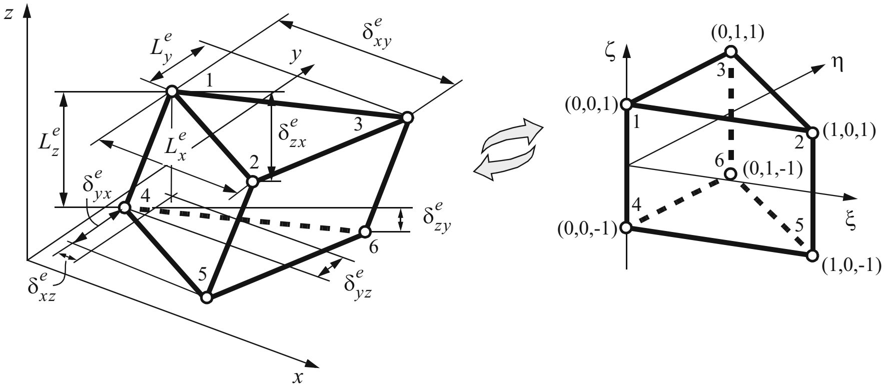


Fig. H.6 Linear 6-node triangular prism with parallel *top* and *bottom* surfaces in the global and local coordinate system

zero. This is valid for the triangular prism with parallel top and bottom surfaces as shown in Fig. H.6, where the following geometric conditions hold:

$$\begin{aligned}
\delta_{xy}^e &= x_3^e - x_1^e = x_6^e - x_4^e \\
\delta_{xz}^e &= x_1^e - x_4^e \\
\delta_{yx}^e &= y_2^e - y_1^e = y_5^e - y_4^e \\
\delta_{yz}^e &= y_1^e - y_4^e \\
\delta_{zx}^e &= z_2^e - z_1^e \;\; = z_5^e - z_4^e \\
\delta_{zy}^e &= z_3^e - z_1^e \;\; = z_6^e - z_4^e \\
L_x^e &= x_2^e - x_1^e = x_5^e - x_4^e \\
L_y^e &= y_3^e - y_1^e = y_6^e - y_4^e \\
L_z^e &= z_1^e - z_4^e
\end{aligned} \tag{H.70}$$

Then, the Jacobain of the triangular prism with parallel top and bottom surfaces

$$\boldsymbol{J}^e = \begin{pmatrix} L_x^e & \delta_{yx}^e & \delta_{zx}^e \\ \delta_{xy}^e & L_y^e & \delta_{zy}^e \\ \delta_{xz}^e & \delta_{yz}^e & L_z^e \end{pmatrix} \tag{H.71}$$

becomes constant. The inverse Jacobian results

$$(\boldsymbol{J}^e)^{-1} = \frac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} (L_y^e L_z^e - \delta_{yz}^e \delta_{zy}^e) & (\delta_{zx}^e \delta_{yz}^e - \delta_{yx}^e L_z^e) & (\delta_{yx}^e \delta_{zy}^e - \delta_{zx}^e L_y^e) \\ (\delta_{xz}^e \delta_{zy}^e - \delta_{xy}^e L_z^e) & (L_x^e L_z^e - \delta_{zx}^e \delta_{xz}^e) & (\delta_{xy}^e \delta_{zx}^e - \delta_{zy}^e L_x^e) \\ (\delta_{xy}^e \delta_{yz}^e - \delta_{xz}^e L_y^e) & (\delta_{yx}^e \delta_{xz}^e - \delta_{yz}^e L_x^e) & (L_x^e L_y^e - \delta_{yx}^e \delta_{xy}^e) \end{pmatrix} \tag{H.72}$$

with the determinant

$$|\boldsymbol{J}^e| = L_x^e(L_y^e L_z^e - \delta_{yz}^e \delta_{zy}^e) + \delta_{xy}^e(\delta_{zx}^e \delta_{yz}^e - \delta_{yx}^e L_z^e) + \delta_{xz}^e(\delta_{yx}^e \delta_{zy}^e - \delta_{zx}^e L_y^e) = 2V^e \tag{H.73}$$

which is twice the volume of the triangular prism. The special case represents a triangular prism which is vertical and has horizontal top and bottom surfaces, so that $\delta_{xz}^e = \delta_{yz}^e = \delta_{zx}^e = \delta_{zy}^e = 0$. Then, it simplifies

$$\begin{aligned}
\boldsymbol{J}^e &= \begin{pmatrix} L_x^e & \delta_{yx}^e & 0 \\ \delta_{xy}^e & L_y^e & 0 \\ 0 & 0 & L_z^e \end{pmatrix} \\
(\boldsymbol{J}^e)^{-1} &= \frac{1}{|\boldsymbol{J}^e|} \begin{pmatrix} L_y^e L_z^e & -\delta_{yx}^e L_z^e & 0 \\ -\delta_{xy}^e L_z^e & L_x^e L_z^e & 0 \\ 0 & 0 & L_x^e L_y^e - \delta_{xy}^e \delta_{yx}^e \end{pmatrix} \\
|\boldsymbol{J}^e| &= (L_x^e L_y^e - \delta_{xy}^e \delta_{yx}^e) L_z^e = 2A^e L_z^e = 2V^e
\end{aligned} \tag{H.74}$$

where $A^e = \frac{1}{2}(L_x^e L_y^e - \delta_{xy}^e \delta_{yx}^e) = \frac{1}{2}[x_1^e(y_2^e - y_3^e) + x_2^e(y_3^e - y_1^e) + x_3^e(y_1^e - y_2^e)]$ is the base area of the triangular prism. In using these relations the element matrices and vectors for the triangular prismatic element can be analytically evaluated in a similar way as done in Sect. H.3 for the tetrahedron.

H.4.4 Linear Pyramidal Element with Parallelogram or Rectangular Base and Oblique Shape

The linear 5-node pyramidal element as described in Appendix G (Table G.3d) has the following shape functions

$$\begin{aligned}
N_1^e &= \tfrac{1}{4}[(1-\xi)(1-\eta) - \zeta + \tfrac{\xi\eta\zeta}{1-\zeta}] \\
N_2^e &= \tfrac{1}{4}[(1+\xi)(1-\eta) - \zeta - \tfrac{\xi\eta\zeta}{1-\zeta}] \\
N_3^e &= \tfrac{1}{4}[(1+\xi)(1+\eta) - \zeta + \tfrac{\xi\eta\zeta}{1-\zeta}] \\
N_4^e &= \tfrac{1}{4}[(1-\xi)(1+\eta) - \zeta - \tfrac{\xi\eta\zeta}{1-\zeta}] \\
N_5^e &= \zeta
\end{aligned} \tag{H.75}$$

and the local derivatives

$$\begin{aligned}
\tfrac{\partial N_1^e}{\partial \xi} &= -\tfrac{1}{4}[(1-\eta) - \tfrac{\eta\zeta}{1-\zeta}], & \tfrac{\partial N_1^e}{\partial \eta} &= -\tfrac{1}{4}[(1-\xi) - \tfrac{\xi\zeta}{1-\zeta}], & \tfrac{\partial N_1^e}{\partial \zeta} &= -\tfrac{1}{4}[1 - \tfrac{\xi\eta}{(1-\zeta)^2}] \\
\tfrac{\partial N_2^e}{\partial \xi} &= \tfrac{1}{4}[(1-\eta) - \tfrac{\eta\zeta}{1-\zeta}], & \tfrac{\partial N_2^e}{\partial \eta} &= -\tfrac{1}{4}[(1+\xi) + \tfrac{\xi\zeta}{1-\zeta}], & \tfrac{\partial N_2^e}{\partial \zeta} &= -\tfrac{1}{4}[1 + \tfrac{\xi\eta}{(1-\zeta)^2}]
\end{aligned}$$

$$\begin{aligned}
\tfrac{\partial N_3^e}{\partial \xi} &= \tfrac{1}{4}[(1+\eta) + \tfrac{\eta\zeta}{1-\zeta}], & \tfrac{\partial N_3^e}{\partial \eta} &= \tfrac{1}{4}[(1+\xi) + \tfrac{\xi\zeta}{1-\zeta}], & \tfrac{\partial N_3^e}{\partial \zeta} &= -\tfrac{1}{4}[1 - \tfrac{\xi\eta}{(1-\zeta)^2}] \\
\tfrac{\partial N_4^e}{\partial \xi} &= -\tfrac{1}{4}[(1+\eta) + \tfrac{\eta\zeta}{1-\zeta}], & \tfrac{\partial N_4^e}{\partial \eta} &= \tfrac{1}{4}[(1-\xi) - \tfrac{\xi\zeta}{1-\zeta}], & \tfrac{\partial N_4^e}{\partial \zeta} &= -\tfrac{1}{4}[1 + \tfrac{\xi\eta}{(1-\zeta)^2}] \\
\tfrac{\partial N_5^e}{\partial \xi} &= 0, & \tfrac{\partial N_5^e}{\partial \eta} &= 0, & \tfrac{\partial N_5^e}{\partial \zeta} &= 1
\end{aligned} \tag{H.76}$$

for $(-1 \leq \xi, \eta \leq 1)$ and $(0 \leq \zeta \leq 1)$. It is for the pyramidal element, cf. (8.71)

$$x = \sum_{I=1}^{5} N_I^e x_I^e, \quad y = \sum_{I=1}^{5} N_I^e y_I^e, \quad z = \sum_{I=1}^{5} N_I^e z_I^e \tag{H.77}$$

where x_I^e, y_I^e, z_I^e $(I = 1, \ldots, 5)$ correspond to the Cartesian coordinates of the vertices (nodes) of the pyramid. Using (H.76) and (H.77) in (8.115) we obtain the Jacobian

$$\boldsymbol{J}^e = \frac{1}{4}\begin{pmatrix} \left(a_{\xi x} + \frac{\eta}{1-\zeta} b_x\right) & \left(a_{\xi y} + \frac{\eta}{1-\zeta} b_y\right) & \left(a_{\xi z} + \frac{\eta}{1-\zeta} b_z\right) \\ \left(a_{\eta x} + \frac{\xi}{1-\zeta} b_x\right) & \left(a_{\eta y} + \frac{\xi}{1-\zeta} b_y\right) & \left(a_{\eta z} + \frac{\xi}{1-\zeta} b_z\right) \\ \left(a_{\zeta x} + \frac{\xi\eta}{(1-\zeta)^2} b_x\right) & \left(a_{\zeta y} + \frac{\xi\eta}{(1-\zeta)^2} b_y\right) & \left(a_{\zeta z} + \frac{\xi\eta}{(1-\zeta)^2} b_z\right) \end{pmatrix} \tag{H.78}$$

where

$$\left.\begin{aligned}
a_{\xi x} &= -x_1^e + x_2^e + x_3^e - x_4^e \\
a_{\xi y} &= -y_1^e + y_2^e + y_3^e - y_4^e \\
a_{\xi z} &= -z_1^e + z_2^e + z_3^e - z_4^e \\
a_{\eta x} &= -x_1^e - x_2^e + x_3^e + x_4^e \\
a_{\eta y} &= -y_1^e - y_2^e + y_3^e + y_4^e \\
a_{\eta z} &= -z_1^e - z_2^e + z_3^e + z_4^e \\
a_{\zeta x} &= -x_1^e - x_2^e - x_3^e - x_4^e + 4x_5^e \\
a_{\zeta y} &= -y_1^e - y_2^e - y_3^e - y_4^e + 4y_5^e \\
a_{\zeta z} &= -z_1^e - z_2^e - z_3^e - z_4^e + 4z_5^e
\end{aligned}\right\} \tag{H.79}$$

$$\left.\begin{aligned}
b_x &= x_1^e - x_2^e + x_3^e - x_4^e \\
b_y &= y_1^e - y_2^e + y_3^e - y_4^e \\
b_z &= z_1^e - z_2^e + z_3^e - z_4^e
\end{aligned}\right\} \tag{H.80}$$

It is obvious from (H.78) that the Jacobian $\boldsymbol{J}^e$ of the pyramid is still a function of the local coordinates ξ, η, ζ. Any analytical evaluation needs that the $\xi, \eta, \zeta-$dependencies in $\boldsymbol{J}^e$ vanish, i.e., all coefficients b_x, b_y and b_z of (H.80) must be zero. This is valid for the pyramid with a parallelogram base and oblique shape as shown in Fig. H.7, where the following geometric conditions hold:

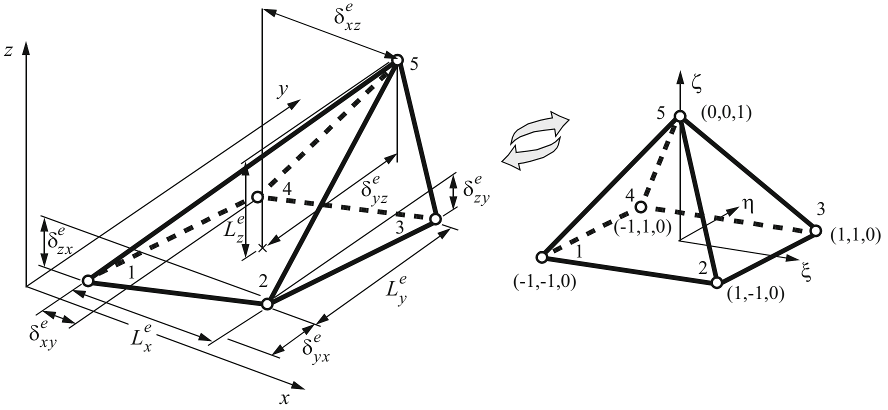


Fig. H.7 Linear 5-node (oblique) pyramid with a parallelogram base in the global and local coordinate system

$$
\begin{aligned}
\delta^e_{xy} &= x^e_4 - x^e_1 = x^e_3 - x^e_2\\
\delta^e_{xz} &= x^e_5 - \tfrac{1}{4}(x^e_1 + x^e_2 + x^e_3 + x^e_4)\\
\delta^e_{yx} &= y^e_2 - y^e_1 = y^e_3 - y^e_4\\
\delta^e_{yz} &= y^e_5 - \tfrac{1}{4}(y^e_1 + y^e_2 + y^e_3 + y^e_4)\\
\delta^e_{zx} &= z^e_2 - z^e_1 = z^e_3 - z^e_4\\
\delta^e_{zy} &= z^e_4 - z^e_1 = z^e_3 - z^e_2\\
L^e_x &= x^e_2 - x^e_1 = x^e_3 - x^e_4\\
L^e_y &= y^e_4 - y^e_1 = y^e_3 - y^e_2\\
L^e_z &= z^e_5 - \tfrac{1}{4}(z^e_1 + z^e_2 + z^e_3 + z^e_4)
\end{aligned}
\tag{H.81}
$$

Then, the Jacobain of the pyramid with a parallelogram base

$$
\boldsymbol{J}^e = \begin{pmatrix} \frac{1}{2}L^e_x & \frac{1}{2}\delta^e_{yx} & \frac{1}{2}\delta^e_{zx}\\ \frac{1}{2}\delta^e_{xy} & \frac{1}{2}L^e_y & \frac{1}{2}\delta^e_{zy}\\ \delta^e_{xz} & \delta^e_{yz} & L^e_z \end{pmatrix}
\tag{H.82}
$$

becomes constant. The inverse Jacobian results

$$
(\boldsymbol{J}^e)^{-1} = \frac{1}{4|\boldsymbol{J}^e|}\begin{pmatrix} 2(L^e_y L^e_z - \delta^e_{yz}\delta^e_{zy}) & 2(\delta^e_{zx}\delta^e_{yz} - \delta^e_{yx}L^e_z) & (\delta^e_{yx}\delta^e_{zy} - \delta^e_{zx}L^e_y)\\ 2(\delta^e_{xz}\delta^e_{zy} - \delta^e_{xy}L^e_z) & 2(L^e_x L^e_z - \delta^e_{zx}\delta^e_{xz}) & (\delta^e_{xy}\delta^e_{zx} - \delta^e_{zy}L^e_x)\\ 2(\delta^e_{xy}\delta^e_{yz} - \delta^e_{xz}L^e_y) & 2(\delta^e_{yx}\delta^e_{xz} - \delta^e_{yz}L^e_x) & (L^e_x L^e_y - \delta^e_{yx}\delta^e_{xy}) \end{pmatrix}
\tag{H.83}
$$

with the determinant

$$|\boldsymbol{J}^e| = \tfrac{1}{4}\big[L_x^e(L_y^e L_z^e - \delta_{yz}^e\delta_{zy}^e) + \delta_{xy}^e(\delta_{zx}^e\delta_{yz}^e - \delta_{yx}^e L_z^e) + \delta_{xz}^e(\delta_{yx}^e\delta_{zy}^e - \delta_{zx}^e L_y^e)\big] = \tfrac{3}{4}V^e \tag{H.84}$$

which is three quarters of the pyramid volume. We note that $\delta_{xz}^e = \delta_{yz}^e = \delta_{zx}^e = \delta_{zy}^e = 0$ for a *right pyramid*, where the apex is aligned directly above the center of the base. The element matrices and vectors can now be analytically evaluated for the pyramid with a parallelogram or rectangular base and possibly oblique shape in a similar way as done in Sect. H.3 for the tetrahedron.

Appendix I
Parameters in Relation to Selected Problem Class, Medium Type and Dimension

The following tables summarize the essential parameters (material/constitutive relationships and BC-related parameters) required for solving the governing flow, mass and heat transport equations as listed in Tables 3.7 and 3.9–3.11 for porous-media problems as well as in Tables 4.5–4.7 for fractured media (discrete feature) problems. They depend on both the problem class (flow, mass transport, heat transport, thermohaline problem), the type of medium (porous medium, fractured medium), the free-surface or variably saturated media formulation and the dimension (3D, 2D (vertical, axisymmetric, horizontal)) of the chosen problem. Generally, the parameters can be functions of space and time, e.g.,

$$\begin{aligned} K_{11} &= K_{11}(\boldsymbol{x}, t) \\ K_{22} &= K_{22}(\boldsymbol{x}, t) \\ K_{33} &= K_{33}(\boldsymbol{x}, t) \\ &\vdots \\ \Phi_h^{\text{out}} &= \Phi_h^{\text{out}}(\boldsymbol{x}, t) \\ &\vdots \end{aligned} \tag{I.1}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

I.1 Flow

Table I.1 Parameters for 3D flow in porous media (with and without mass transport)

Item	Symbol	Unit	Default	Reference(s)
Axis-parallel anisotropy:				
Conductivity [Kxx]	K_{11}	10^{-4} m s^{-1}	1	Sect. 7.4.1
Conductivity [Kyy]	K_{22}	10^{-4} m s^{-1}	1	
Conductivity [Kzz]	K_{33}	10^{-4} m s^{-1}	1	
General and shaped-derived anisotropy (optional):				
Conductivity [K1m]	K_1^m	10^{-4} m s^{-1}	1	Sects. 7.3.1 and 7.3.2
Conductivity [K2m]	K_2^m	10^{-4} m s^{-1}	1	
Conductivity [K3m]	K_3^m	10^{-4} m s^{-1}	1	
Eulerian angles only for general anisotropy:				
ϕ	ϕ	°	0	Sect. 7.3.1
θ	θ	°	0	
ψ	ψ	°	0	
In(+)/out(−)flow on top/bottom	P	10^{-4} m d^{-1}	0	(9.4)
Density ratio	α_k	10^{-4}	0	(3.199) and (3.275)
Specific yield	ε_e	1	0.2	(3.296) and (9.4)
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25) and (9.2)
Source(+)/sink(−)	Q_h	10^{-4} d^{-1}	0	(9.2)
In-transfer coefficient	Φ_h^{in}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Out-transfer coefficient	Φ_h^{out}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Unsaturated properties	*as listed in Table I.10*			

Table I.2 Parameters for 3D flow in porous media (at heat transport)

Item	Symbol	Unit	Default	Reference(s)
Axis-parallel anisotropy:				
Conductivity [Kxx]	K_{11}	10^{-4} m s^{-1}	1	Sect. 7.4.1
Conductivity [Kyy]	K_{22}	10^{-4} m s^{-1}	1	
Conductivity [Kzz]	K_{33}	10^{-4} m s^{-1}	1	
General and shaped-derived anisotropy (optional):				
Conductivity [K1m]	K_1^m	10^{-4} m s^{-1}	1	Sects. 7.3.1 and 7.3.2
Conductivity [K2m]	K_2^m	10^{-4} m s^{-1}	1	
Conductivity [K3m]	K_3^m	10^{-4} m s^{-1}	1	
Eulerian angles only for general anisotropy:				
ϕ	ϕ	°	0	Sect. 7.3.1
θ	θ	°	0	
ψ	ψ	°	0	
In(+)/out(−)flow on top/bottom	P	10^{-4} m d^{-1}	0	(9.4)

(continued)

Table I.2 (continued)

Item	Symbol	Unit	Default	Reference(s)
Constant thermal expansion coefficient:				
Expansion coefficient (constant)	β	10^{-4} K^{-1}	0	(3.199)
Variable thermal expansion (optional):				
Expansion coefficient (variable)	$\beta(T)$	10^{-4} K^{-1}	0	(C.4) and (C.8)
Specific yield	ε_e	1	0.2	(3.296) and (9.4)
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25) and (9.2)
Source(+)/sink(−)	Q_h	10^{-4} d^{-1}	0	(9.2)
In-transfer coefficient	Φ_h^{in}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Out-transfer coefficient	Φ_h^{out}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Unsaturated properties	*as listed in Table I.10*			

Table I.3 Parameters for 3D flow in porous media (at thermohaline transport)

Item	Symbol	Unit	Default	Reference(s)
Axis-parallel anisotropy:				
Conductivity [Kxx]	K_{11}	10^{-4} m s^{-1}	1	Sect. 7.4.1
Conductivity [Kyy]	K_{22}	10^{-4} m s^{-1}	1	
Conductivity [Kzz]	K_{33}	10^{-4} m s^{-1}	1	
General and shaped-derived anisotropy (optional):				
Conductivity [K1m]	K_1^m	10^{-4} m s^{-1}	1	Sects. 7.3.1 and 7.3.2
Conductivity [K2m]	K_2^m	10^{-4} m s^{-1}	1	
Conductivity [K3m]	K_3^m	10^{-4} m s^{-1}	1	
Eulerian angles only for general anisotropy:				
ϕ	ϕ	°	0	Sect. 7.3.1
θ	θ	°	0	
ψ	ψ	°	0	
In(+)/out(-)flow on top/bottom	P	10^{-4} m d^{-1}	0	(9.4)
Density ratio	α_k	10^{-4}	0	(3.199) and (3.275)
Constant thermal expansion coefficient:				
Expansion coefficient (constant)	β	10^{-4} K^{-1}	0	(3.199)
Variable thermal expansion (optional):				
Expansion coefficient (variable)	$\beta(T)$	10^{-4} K^{-1}	0	(C.4) and (C.8)
Specific yield	ε_e	1	0.2	(3.296) and (9.4)
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25) and (9.2)
Source(+)/sink(−)	Q_h	10^{-4} d^{-1}	0	(9.2)
In-transfer coefficient	Φ_h^{in}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Out-transfer coefficient	Φ_h^{out}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Unsaturated properties	*as listed in Table I.10*			

Table I.4 Parameters for vertical or axisymmetric 2D flow in porous media (with and without mass transport)

Item	Symbol	Unit	Default	Reference(s)
Conductivity [Kmax]	$K_{\max}$	10^{-4} m s^{-1}	1	Sect. 7.2
Anisotropy factor [Kmin/Kmax]	Ξ_{aniso}	1	1	(7.9)
Angle from + x-axis to Kmax	ϕ	°	0	Fig. 7.2
Density ratio	α_k	10^{-4}	0	(3.199) and (3.275)
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25), (9.2)
Source(+)/sink(−)	Q_h	10^{-4} d^{-1}	0	(9.2)
In-transfer coefficient	Φ_h^{in}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Out-transfer coefficient	Φ_h^{out}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Unsaturated properties	*as listed in Table I.10*			

Table I.5 Parameters for vertical or axisymmetric 2D flow in porous media (at heat transport)

Item	Symbol	Unit	Default	Reference(s)
Conductivity [Kmax]	$K_{\max}$	10^{-4} m s^{-1}	1	Sect. 7.2
Anisotropy factor [Kmin/Kmax]	Ξ_{aniso}	1	1	(7.9)
Angle from + x-axis to Kmax	ϕ	°	0	Fig. 7.2
Constant thermal expansion coefficient:				
Expansion coefficient (constant)	β	10^{-4} K^{-1}	0	(3.199)
Variable thermal expansion (optional):				
Expansion coefficient (variable)	$\beta(T)$	10^{-4} K^{-1}	0	(C.4) and (C.8)
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25) and (9.2)
Source(+)/sink(−)	Q_h	10^{-4} d^{-1}	0	(9.2)
In-transfer coefficient	Φ_h^{in}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Out-transfer coefficient	Φ_h^{out}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Unsaturated properties	*as listed in Table I.10*			

Table I.6 Parameters for vertical or axisymmetric 2D flow in porous media (at thermohaline transport)

Item	Symbol	Unit	Default	Reference(s)
Conductivity [Kmax]	$K_{\max}$	10^{-4} m s^{-1}	1	Sect. 7.2
Anisotropy factor [Kmin/Kmax]	Ξ_{aniso}	1	1	(7.9)
Angle from + x-axis to Kmax	ϕ	°	0	Fig. 7.2
Density ratio	α_k	10^{-4}	0	(3.199) and (3.275)
Constant thermal expansion coefficient:				
Expansion coefficient (constant)	β	10^{-4} K^{-1}	0	(3.199)
Variable thermal expansion (optional):				
Expansion coefficient (variable)	$\beta(T)$	10^{-4} K^{-1}	0	(C.4) and (C.8)
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25) and (9.2)
Source(+)/sink(−)	Q_h	10^{-4} d^{-1}	0	(9.2)
In-transfer coefficient	Φ_h^{in}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Out-transfer coefficient	Φ_h^{out}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Unsaturated properties	*as listed in Table I.10*			

Table I.7 Parameters for horizontal 2D flow in porous media (confined aquifer)

Item	Symbol	Unit	Default	Reference(s)
Transmissivity [Tmax]	$T_{\max}$	10^{-4} m^2 s^{-1}	1	Sect. 7.2, (3.302)
Anisotropy factor [Tmin/Tmax]	Ξ_{aniso}	1	1	(7.9)
Angle from + x-axis to Tmax	ϕ	°	0	Fig. 7.2
Specific storage coefficient	$\bar{S}_o$	1	10^{-4}	(3.299) and (9.11)
Source(+)/sink(−)	$\bar{Q}_h$	10^{-4} m d^{-1}	0	(9.12)
In-transfer coefficient	$\bar{\Phi}_h^{\text{in}}$	10^{-4} m d^{-1}	0	(6.7) and (6.9)
Out-transfer coefficient	$\bar{\Phi}_h^{\text{out}}$	10^{-4} m d^{-1}	0	(6.7) and (6.9)

Table I.8 Parameters for horizontal 2D flow in porous media (unconfined aquifer)

Item	Symbol	Unit	Default	Reference(s)
Conductivity [Kmax]	$K_{\max}$	10^{-4} m s^{-1}	1	Sect. 7.2
Anisotropy factor [Kmin/Kmax]	Ξ_{aniso}	1	1	(7.9)
Angle from + x-axis to Kmax	ϕ	°	0	Fig. 7.2
Aquifer bottom elevation	f^B	m	0	(3.283) and (9.7)
Aquifer top elevation	f^T	m	10^3	(3.283) and (9.31)
Specific yield	ε_e	1	0.2	(3.296) and (9.8)
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25) and (9.8)
Source(+)/sink(−)	$\bar{Q}_h$	10^{-4} m d^{-1}	0	(9.8)
In-transfer coefficient[a]	Φ_h^{in}	10^{-4} d^{-1}	0	(6.7) and (6.8)
Out-transfer coefficient[b]	Φ_h^{out}	10^{-4} d^{-1}	0	(6.7) and (6.8)

[a] For integral third kind BC (Sect. 6.5.4) depth-integrated $\bar{\Phi}_h^{\text{in}}$ [10^{-4} md^{-1}] (6.9) is used

[b] For integral third kind BC (Sect. 6.5.4) depth-integrated $\bar{\Phi}_h^{\text{out}}$ [10^{-4} md^{-1}] (6.9) is used

Table I.9 Parameters for flow in fractured media (discrete features)

Item	Symbol	Unit	Default	Reference(s)
1D fracture type:				
Cross-sectional area	A	m^2	1	Table 4.8
2D fracture type:				
Thickness	B	m	10^{-3}	Table 4.8
Darcy flux law:				
Conductivity	K	10^{-4} m s^{-1}	1	(4.38), Table 4.5
Hagen-Poiseuille flux law:				
Hydraulic aperture	b	m	10^{-3}	Table 4.9
Manning-Strickler flux law:				
Roughness coefficient	M	m$^{1/3}$ s^{-1}	30	Tables 4.9 and 4.4
Specific yield	ε_e	1	1	(4.29), Table 4.5
Specific storage coefficient	S_o	m^{-1}	10^{-4}	(4.25), Table 4.5
Source(+)/sink(−)	Q	10^{-4} d^{-1}	0	Table 4.5
Density ratio	α_k	10^{-4}	0	(4.63) and (3.265)

(continued)

Table I.9 (continued)

Item	Symbol	Unit	Default	Reference(s)
Constant thermal expansion coefficient:				
Expansion coefficient (constant)	β	$10^{-4}\ \mathrm{K}^{-1}$	0	(4.63) and (3.265)
Variable thermal expansion (optional):				
Expansion coefficient (variable)	$\beta(T)$	$10^{-4}\ \mathrm{K}^{-1}$	0	(4.63) and (C.4)
In-transfer coefficient	Φ_h^{in}	$10^{-4}\ \mathrm{d}^{-1}$	0	(6.7) and (6.8)
Out-transfer coefficient	Φ_h^{out}	$10^{-4}\ \mathrm{d}^{-1}$	0	(6.7) and (6.8)

Table I.10 Parameters for unsaturated porous media (analytical parametric models summarized in Table D.1) (Spline approximations alternatively exist to input experimental sample points for the parametric curves of saturation s and relative permeability k_r, see Sect. D.4 of Appendix D)

Item	Symbol	Unit	Default	Reference(s)
Porosity	ε	1	0.3	(3.219)
Residual saturation	s_r	1	0.0025	(D.1)
Maximum saturation	s_s	1	1.0	(D.1)
van Genuchten-Mualem parametric model:				
Fitting coefficient	α	m^{-1}	4.1	(D.3)
Fitting exponent	n	1	1.964	(D.3)
Modified van Genuchten parametric model:				
Fitting coefficient	α	m^{-1}	4.1	(D.3)
Fitting exponent	n	1	1.964	(D.3)
Fitting exponent	m	1	0.491	(D.3)
Fitting exponent	δ	1	3.4	(D.32)
Brooks-Corey parametric model:				
Fitting coefficient	α	m^{-1}	4.1	(D.7)
Fitting exponent	n	1	1.964	(D.7)
Fitting exponent	δ	1	3.4	(D.29)
Haverkamp parametric model:				
Fitting coefficient	α	1	4.1	(D.11)
Fitting exponent	β	1	1.964	(D.11)
Fitting coefficient	A	1	3.4	(D.35)
Fitting exponent	B	1	0.491	(D.35)
Exponential parametric model:				
Sorptive number	α	m^{-1}	3.4	(D.15) and (D.38)
Air-entry pressure head	ψ_a	m	0	(D.15) and (D.38)
Linear parametric model:				
Fringe pressure head	ψ_c	m	−4.1	(D.19) and (D.41)
Air-entry pressure head	ψ_a	m	0	(D.19) and (D.41)
Hysteresis[a]				

[a] Parametric models under hysteretic conditions require a double dataset for drying and wetting curves

I.2 Mass Transport

Table I.11 Parameters for 3D, vertical or axisymmetric mass transport of species k at liquid phase l in porous media

Item	Symbol	Unit	Default	Reference(s)
Porosity	ε	1	0.3	(3.219)
Henry sorption:				
Sorption coefficient	κ_k	1	0	(5.65), Table 3.8
Freundlich sorption:				
Sorption coefficient	$b_k^{\dagger}$	$(\mathrm{mg\,l^{-1}})^{1-b_k^{\ddagger}}$	0	(5.65), Table 3.8
Sorption exponent	$b_k^{\ddagger}$	1	0	(5.65), Table 3.8
Langmuir sorption:				
Numerator sorption coefficient	$k_k^{\dagger}$	1	0	(5.65), Table 3.8
Denominator sorption coefficient	$k_k^{\ddagger}$	$\mathrm{l(mg)^{-1}}$	0	(5.65), Table 3.8
Molecular diffusion	D_k	$10^{-9}\,\mathrm{m^2\,s^{-1}}$	1	(3.184), Table 3.7
Longitudinal dispersivity	β_L	m	5	(3.184), Table 3.7
Transverse dispersivity	β_T	m	0.5	(3.184), Table 3.7
Nonlinear (non-Fickian) dispersion:				
HC dispersion coefficient	$\Im_H$	$\mathrm{m^2\,d\,g^{-1}}$	0	(3.272), Table 3.7
Chemical reactions for single-species solute transport:				
First-order decay:				
Decay rate	ϑ_k	$10^{-4}\,\mathrm{s^{-1}}$	0	(5.73), Table 3.7
Michaelis-Menten:				
Maximum growth rate	v_m	$10^{-4}\,\mathrm{mg\,l^{-1}s^{-1}}$	0	(5.90)
Half-saturation constant	K_m	$\mathrm{mg\,l^{-1}}$	0	(5.90)
Chemical reactions for multispecies mass transport:				
Rate constant	k_k	$10^{-4}\,\mathrm{s^{-1}}$	0	Sect. 5.5
Σ	⇒ Reaction kinetics editor			
Source(+)/sink(−)	Q_k	$\mathrm{g\,m^{-3}\,d^{-1}}$	0	
In-transfer coefficient[a]	Φ_{kC}^{in}	$\mathrm{m\,d^{-1}}$	0	(6.22) and (6.24)
Out-transfer coefficient[b]	Φ_{kC}^{out}	$\mathrm{m\,d^{-1}}$	0	(6.22) and (6.24)

[a] For the divergence form $\Phi_{kC}^{\dagger\,\mathrm{in}}$ is input

[b] For the divergence form $\Phi_{kC}^{\dagger\,\mathrm{out}}$ is input

Table I.12 Parameters for 2D horizontal mass transport of species k at liquid phase l in porous media (unconfined and confined aquifer)

Item	Symbol	Unit	Default	Reference(s)
Confined aquifer:				
Aquifer thickness	B	m	1.0	(3.283)
Porosity	ε	1	0.3	(3.219)
Henry sorption:				
Sorption coefficient	κ_k	1	0	(5.65), Table 3.8
Freundlich sorption:				
Sorption coefficient	$b_k^\dagger$	$(\text{mg}\,\text{l}^{-1})^{1-b_k^\ddagger}$	0	(5.65), Table 3.8
Sorption exponent	$b_k^\ddagger$	1	0	(5.65), Table 3.8
Langmuir sorption:				
Numerator sorption coefficient	$k_k^\dagger$	1	0	(5.65), Table 3.8
Denominator sorption coefficient	$k_k^\ddagger$	$\text{l}(\text{mg})^{-1}$	0	(5.65), Table 3.8
Molecular diffusion	D_k	$10^{-9}\,\text{m}^2\,\text{s}^{-1}$	1	(3.184), Table 3.10
Longitudinal dispersivity	β_L	m	5	(3.184), Table 3.10
Transverse dispersivity	β_T	m	0.5	(3.184), Table 3.10
Nonlinear (non-Fickian) dispersion:				
HC dispersion coefficient	$\bar{\Im}_H$	$\text{m}\,\text{d}\,\text{g}^{-1}$	0	(3.272), Table 3.10
Chemical reactions for single-species solute transport:				
First-order decay:				
Decay rate	ϑ_k	$10^{-4}\,\text{s}^{-1}$	0	(5.73), Table 3.10
Michaelis-Menten:				
Maximum growth rate	v_m	$10^{-4}\,\text{mg}\,\text{l}^{-1}\text{s}^{-1}$	0	(5.90)
Half-saturation constant	K_m	$\text{mg}\,\text{l}^{-1}$	0	(5.90)
Chemical reactions for multispecies mass transport:				
Rate constant	k_k	$10^{-4}\,\text{s}^{-1}$	0	Sect. 5.5
Σ	⇒ Reaction kinetics editor			
Source(+)/sink(−)	$\bar{Q}_k$	$\text{gm}^{-2}\,\text{d}^{-1}$	0	
In-transfer coefficient[a]	Φ_{kC}^{in}	$\text{m}\,\text{d}^{-1}$	0	(6.22) and (6.24)
Out-transfer coefficient[b]	Φ_{kC}^{out}	$\text{m}\,\text{d}^{-1}$	0	(6.22) and (6.24)

[a] For the divergence form $\Phi_{kC}^{\dagger\,\text{in}}$ is input. For confined conditions or integral third kind BC (Sect. 6.5.4) $\bar{\Phi}_{kC}^{\text{in}}$ $[\text{m}^2\text{d}^{-1}]$ (6.23) and (6.25) is used

[b] For the divergence form $\Phi_{kC}^{\dagger\,\text{out}}$ is input. For confined conditions or integral third kind BC (Sect. 6.5.4) $\bar{\Phi}_{kC}^{\text{out}}$ $[\text{m}^2\text{d}^{-1}]$ (6.23) and (6.25) is used

Table I.13 Parameters for 3D, vertical or axisymmetric mass transport of species k at solid phase s in porous media

Item	Symbol	Unit	Default	Reference(s)
Solid volume fraction	ε_s	1	0.7	(3.219)
First-order decay:				
Decay rate	ϑ_k	$10^{-4}\,\text{s}^{-1}$	0	(5.73), Table 3.7

(continued)

Table I.13 (continued)

Item	Symbol	Unit	Default	Reference(s)
Chemical reactions for multispecies mass transport:				
Rate constant	k_k	$10^{-4}\ \mathrm{s}^{-1}$	0	Sect. 5.5
Σ	⇒ Reaction kinetics editor			
Source(+)/sink(−)	Q_k	$\mathrm{gm}^{-3}\ \mathrm{d}^{-1}$	0	

Table I.14 Parameters for 2D horizontal mass transport of species k at solid phase s in porous media (unconfined and confined aquifer)

Item	Symbol	Unit	Default	Reference(s)
Confined aquifer:				
Aquifer thickness	B	m	1.0	(3.283)
Solid volume fraction	ε_s	1	0.7	(3.219)
First-order decay:				
Decay rate	ϑ_k	$10^{-4}\ \mathrm{s}^{-1}$	0	(5.73), Table 3.7
Chemical reactions for multispecies mass transport:				
Rate constant	k_k	$10^{-4}\ \mathrm{s}^{-1}$	0	Sect. 5.5
Σ	⇒ Reaction kinetics editor			
Source(+)/sink(−)	Q_k	$\mathrm{gm}^{-3}\ \mathrm{d}^{-1}$	0	

Table I.15 Parameters for mass transport of species k at liquid phase l in fractured media (discrete features)

Item	Symbol	Unit	Default	Reference(s)
1D fracture type:				
Cross-sectional area	A	m^2	1	Table 4.8
2D fracture type:				
Thickness	B	m	10^{-3}	Table 4.8
Darcy flux law only:				
Porosity	ε	1	1.0	(4.7)
Henry sorption coefficient	κ_k	1	0	(5.65), Table 3.8
Molecular diffusion	D_k	$10^{-9}\ \mathrm{m}^2\ \mathrm{s}^{-1}$	1	(4.67), Table 4.6
Longitudinal dispersivity	β_L	m	5	(4.68), Table 4.6
Transverse dispersivity	β_T	m	0.5	(4.68), Table 4.6
Chemical reactions for single-species solute transport:				
Decay rate	ϑ_k	$10^{-4}\ \mathrm{s}^{-1}$	0	(4.66), Table 4.6
Chemical reactions for multispecies mass transport:				
Rate constant	k_k	$10^{-4}\ \mathrm{s}^{-1}$	0	Sect. 5.5
Σ	⇒ Reaction kinetics editor			
Source(+)/sink(−)	Q_k	$\mathrm{gm}^{-3}\ \mathrm{d}^{-1}$	0	
In-transfer coefficient[a]	Φ_{kC}^{in}	$\mathrm{m\,d}^{-1}$	0	(6.22) and (6.24)
Out-transfer coefficient[b]	Φ_{kC}^{out}	$\mathrm{m\,d}^{-1}$	0	(6.22) and (6.24)

[a] For the divergence form $\Phi_{kC}^{\dagger\,\mathrm{in}}$ is input

[b] For the divergence form $\Phi_{kC}^{\dagger\,\mathrm{out}}$ is input

I.3 Heat Transport

Table I.16 Parameters for 3D, vertical or axisymmetric heat transport in porous media

Item	Symbol	Unit	Default	Reference(s)
Porosity	ε	1	0.3	(3.219)
Volumetric heat capacity of fluid	ρc	$10^6\,\mathrm{J\,m^{-3}\,K^{-1}}$	4.2[a]	(3.208), Table 3.7
Volumetric heat capacity of solid	$\rho^s c^s$	$10^6\,\mathrm{J\,m^{-3}\,K^{-1}}$	2.52[b]	(3.208), Table 3.7
Heat conductivity of fluid	Λ	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$	0.65	(3.172), Table 3.7
Heat conductivity of solid	Λ^s	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$	3	(3.172), Table 3.7
3D problems only:				
Anisotropy factor $[\Lambda^s_{zz}/\Lambda^s_{xx,yy}]$	$\Xi^{\Lambda}_{\mathrm{aniso}}$	1	1	(7.26)
Longitudinal dispersivity	β_L	m	5	(3.238), Table 3.7
Transverse dispersivity	β_T	m	0.5	(3.238), Table 3.7
Source(+)/sink(−) of fluid	$\rho H^\star$	$\mathrm{J\,m^{-3}\,d^{-1}}$	0	(13.3), Table 3.7
Source(+)/sink(−) of solid	$\rho^s H^\star_s$	$\mathrm{J\,m^{-3}\,d^{-1}}$	0	(13.3), Table 3.7
In-transfer coefficient[c]	Φ^{in}_T	$\mathrm{J\,m^{-2}\,d^{-1}\,K^{-1}}$	0	(6.40) and (6.42)
Out-transfer coefficient[d]	Φ^{out}_T	$\mathrm{J\,m^{-2}\,d^{-1}\,K^{-1}}$	0	(6.40) and (6.42)

[a] $\rho = 1{,}000\,\mathrm{kg\,m^{-3}}$, $c = 4{,}200\,\mathrm{J\,kg^{-1}\,K^{-1}}$

[b] $\rho^s = 2{,}650\,\mathrm{kg\,m^{-3}}$, $c^s = 950\,\mathrm{J\,kg^{-1}\,K^{-1}}$

[c] For the divergence form $\Phi^{\dagger\,\mathrm{in}}_T$ is input

[d] For the divergence form $\Phi^{\dagger\,\mathrm{out}}_T$ is input

Table I.17 Parameters for 2D horizontal heat transport in porous media (unconfined and confined aquifer)

Item	Symbol	Unit	Default	Reference(s)
Confined aquifer:				
Aquifer thickness	B	m	1.0	(3.283)
Porosity	ε	1	0.3	(3.219)
Volumetric heat capacity of fluid	ρc	$10^6\,\mathrm{J\,m^{-3}\,K^{-1}}$	4.2[a]	(3.208), Table 3.10
Volumetric heat capacity of solid	$\rho^s c^s$	$10^6\,\mathrm{J\,m^{-3}\,K^{-1}}$	2.52[b]	(3.208), Table 3.10
Heat conductivity of fluid	Λ	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$	0.65	(3.172), Table 3.10
Heat conductivity of solid	Λ^s	$\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$	3	(3.172), Table 3.10
Longitudinal dispersivity	β_L	m	5	(3.238), Table 3.10
Transverse dispersivity	β_T	m	0.5	(3.238), Table 3.10
Source(+)/sink(−) of fluid	$B\rho H^\star$	$\mathrm{J\,m^{-2}\,d^{-1}}$	0	Table 3.10
Source(+)/sink(−) of solid	$B\rho^s H^\star_s$	$\mathrm{J\,m^{-2}\,d^{-1}}$	0	Table 3.10
In-transfer coefficient[c]	Φ^{in}_T	$\mathrm{J\,m^{-2}\,d^{-1}\,K^{-1}}$	0	(6.40) and (6.42)
Out-transfer coefficient[d]	Φ^{out}_T	$\mathrm{J\,m^{-2}\,d^{-1}\,K^{-1}}$	0	(6.40) and (6.42)

[a] $\rho = 1{,}000\,\mathrm{kg\,m^{-3}}$, $c = 4{,}200\,\mathrm{J\,kg^{-1}\,K^{-1}}$

[b] $\rho^s = 2{,}650\,\mathrm{kg\,m^{-3}}$, $c^s = 950\,\mathrm{J\,kg^{-1}\,K^{-1}}$

[c] For the divergence form $\Phi^{\dagger\,\mathrm{in}}_T$ is input. For confined conditions or integral third kind BC (Sect. 6.5.4) depth-integrated $\bar{\Phi}^{\mathrm{in}}_T$ $[\mathrm{J\,m^{-1}\,d^{-1}\,K^{-1}}]$ (6.41) and (6.43) is used

[d] For the divergence form $\Phi^{\dagger\,\mathrm{out}}_T$ is input. For confined conditions or integral third kind BC (Sect. 6.5.4) depth-integrated $\bar{\Phi}^{\mathrm{out}}_T$ $[\mathrm{J\,m^{-1}\,d^{-1}\,K^{-1}}]$ (6.41) and (6.43) is used

Table I.18 Parameters for heat transport in fractured media (discrete features)

Item	Symbol	Unit	Default	Reference(s)
1D fracture type:				
Cross-sectional area	A	m^2	1	Table 4.8
2D fracture type:				
Thickness	B	m	10^{-3}	Table 4.8
Darcy flux law only:				
Porosity	ε	1	1.0	(4.7)
Volumetric heat capacity of fluid	ρc	$10^6\,J\,m^{-3}\,K^{-1}$	4.2[a]	(4.75), Table 4.7
Darcy flux law only:				
Volumetric heat capacity of solid	$\rho^s c^s$	$10^6\,J\,m^{-3}\,K^{-1}$	2.52[b]	(4.75), Table 4.7
Heat conductivity of fluid	Λ	$J\,m^{-1}\,s^{-1}\,K^{-1}$	0.65	(4.76), Table 4.7
Darcy flux law only:				
Heat conductivity of solid	Λ^s	$J\,m^{-1}\,s^{-1}\,K^{-1}$	3	(4.76), Table 4.7
Longitudinal dispersivity	β_L	m	5	(4.68), Table 4.7
Transverse dispersivity	β_T	m	0.5	(4.68), Table 4.7
Source(+)/sink(−) of fluid	ρH	$J\,m^{-3}\,d^{-1}$	0	(4.74), Table 4.7
Darcy flux law only:				
Source(+)/sink(−) of solid	$\rho^s H_s$	$J\,m^{-3}\,d^{-1}$	0	(4.74), Table 4.7
In-transfer coefficient[c]	Φ_T^{in}	$J\,m^{-2}\,d^{-1}\,K^{-1}$	0	(6.40) and (6.42)
Out-transfer coefficient[d]	Φ_T^{out}	$J\,m^{-2}\,d^{-1}\,K^{-1}$	0	(6.40) and (6.42)

[a] $\rho = 1{,}000\,kg\,m^{-3}$, $c = 4{,}200\,J\,kg^{-1}\,K^{-1}$

[b] $\rho^s = 2{,}650\,kg\,m^{-3}$, $c^s = 950\,J\,kg^{-1}\,K^{-1}$

[c] For the divergence form $\Phi_T^{\dagger\,in}$ is input

[d] For the divergence form $\Phi_T^{\dagger\,out}$ is input

Appendix J
Elements of PVST for Solving the Mixed $\psi - s$–Based Form of Richards' Equation

J.1 Jacobian J^{ψ} for the Pressure Head ψ as Primary Variable

The derivative of the residual (10.64) with respect to the pressure head $\boldsymbol{\psi}_{n+1}^{\tau}$ at the new time plane $n+1$ and the current iterate τ yields the following expression $(i, j, l = 1, \ldots, N_{\mathrm{P}})$:

$$\begin{aligned} J_{ij}^{\psi}(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau}) &= \frac{\partial R_{i,n+1}(\boldsymbol{\psi}_{n+1}^{\tau}, \boldsymbol{s}_{n+1}^{\tau})}{\partial \psi_{j,n+1}^{\tau}} \\ &= J_{ij}^{\psi 1} + J_{ij}^{\psi 2} + J_{ij}^{\psi 3} + J_{ij}^{\psi 4} - J_{ij}^{\psi 5} \\ &= \frac{O_{ij}(\boldsymbol{s}_{n+1}^{\tau})}{\theta \Delta t_n} + D_{ij}(\boldsymbol{\psi}_{n+1}^{\tau}) + \\ &\quad \frac{B_{il}}{\theta \Delta t_n} \frac{\partial s_{l,n+1}^{\tau}}{\partial \psi_{j,n+1}^{\tau}} + \\ &\quad \frac{\partial O_{il}(\boldsymbol{\psi}_{n+1}^{\tau})}{\partial \psi_{j,n+1}^{\tau}} \left(\frac{\psi_{l,n+1}^{\tau} - \psi_{l,n}}{\theta \Delta t_n} - \left(\tfrac{1}{\theta} - 1\right) \dot{\psi}_{l,n} \right) + \\ &\quad \psi_{l,n+1}^{\tau} \frac{\partial D_{il}(\boldsymbol{s}_{n+1}^{\tau})}{\partial \psi_{j,n+1}^{\tau}} - \\ &\quad \frac{\partial F_i(\boldsymbol{s}_{n+1}^{\tau})}{\partial \psi_{j,n+1}^{\tau}} \end{aligned} \qquad \text{(J.1)}$$

where the matrices $\boldsymbol{O}$, $\boldsymbol{B}$, $\boldsymbol{D}$ and the RHS-vector $\boldsymbol{F}$ are defined in (10.61). The partial Jacobians in (J.1) are obtained as follows

$$J_{ij}^{\psi 1} = \frac{O_{ij}(\boldsymbol{s}_{n+1}^{\tau})}{\theta \Delta t_n} + D_{ij}(\boldsymbol{\psi}_{n+1}^{\tau}) \qquad \text{(J.2)}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

$$J_{ij}^{\psi 2} = \sum_e \int_{\Omega^e} N_i \, \varepsilon^e C_{j,n+1}^{\tau} \delta_{ij} \frac{1}{\theta \Delta t_n} d\Omega^e \qquad \text{(J.3)}$$
$$\text{no summation over } i \text{ and } j$$

$$J_{ij}^{\psi 3} = \sum_e \int_{\Omega^e} N_i \, S_o^e C_{j,n+1}^{\tau} \delta_{ij} \left(\frac{\psi_{j,n+1}^{\tau} - \psi_{j,n}}{\theta \Delta t_n} - \big(\tfrac{1}{\theta} - 1\big) \dot{\psi}_{j,n} \right) d\Omega^e \qquad \text{(J.4)}$$
$$\text{no summation over } i \text{ and } j$$

$$J_{ij}^{\psi 4} = \sum_e \int_{\Omega^e} \nabla N_i \cdot \big(\boldsymbol{K}^e f_{\mu}^e N_j G_{j,n+1}^{\tau} \cdot \nabla N_l \psi_{l,n+1}^{\tau} \big) d\Omega^e \qquad \text{(J.5)}$$
$$\text{no summation over } j$$

$$J_{ij}^{\psi 5} = -\sum_e \Bigg(\int_{\Omega^e} \nabla N_i \cdot \big(\boldsymbol{K}^e f_{\mu}^e N_j (1 + \chi^e) G_{j,n+1}^{\tau} \cdot \boldsymbol{e} \big) d\Omega^e + \int_{\Gamma_N^{\nabla e}} N_i N_j G_{j,n+1}^{\tau} q_h^{\nabla e} \, d\Gamma^e \Bigg) \qquad \text{(J.6)}$$
$$\text{no summation over } j$$

with

$$C_{j,n+1}^{\tau} = \frac{\partial s(\psi_{j,n+1}^{\tau})}{\partial \psi_{j,n+1}^{\tau}} \qquad \text{(J.7)}$$

and

$$G_{j,n+1}^{\tau} = \frac{\partial k_r(\psi_{j,n+1}^{\tau})}{\partial \psi_{j,n+1}^{\tau}} \qquad \text{(J.8)}$$

The derivatives $C_{j,n+1}^{\tau}$ and $G_{j,n+1}^{\tau}$ are given functions which can be evaluated either analytically from the parametric models summarized in Appendix D or numerically from chord slope approximations given in Sect. J.3 for the known variables s and ψ at the iterate τ, the global node j and the time plane $n + 1$. Here, $\boldsymbol{C}_{n+1}^{\tau}$ is the moisture capacity function known from the standard unsaturated flow modeling.

J.2 Jacobian J^s for the Saturation s as Primary Variable

The derivative of the residual (10.64) with respect to the saturation $\boldsymbol{s}_{n+1}^{\tau}$ at the new time plane $n + 1$ and the current iterate τ yields the following expression $(i, j, l = 1, \ldots, N_{\mathrm{P}})$:

$$
\begin{aligned}
J^s_{ij}(\boldsymbol{\psi}^\tau_{n+1}, \boldsymbol{s}^\tau_{n+1}) &= \frac{\partial R_{i,n+1}(\boldsymbol{\psi}^\tau_{n+1}, \boldsymbol{s}^\tau_{n+1})}{\partial s^\tau_{j,n+1}} \\
&= J^{s1}_{ij} + J^{s2}_{ij} + J^{s3}_{ij} + J^{s4}_{ij} - J^{s5}_{ij} \\
&= \left(\frac{O_{il}(\boldsymbol{s}^\tau_{n+1})}{\theta \Delta t_n} + D_{il}(\boldsymbol{\psi}^\tau_{n+1}) \right) \frac{\partial \psi^\tau_{l,n+1}}{\partial s^\tau_{j,n+1}} + \\
&\quad \frac{B_{ij}}{\theta \Delta t_n} + \\
&\quad \frac{\partial O_{il}(\boldsymbol{\psi}^\tau_{n+1})}{\partial s^\tau_{j,n+1}} \left(\frac{\psi^\tau_{l,n+1} - \psi_{l,n}}{\theta \Delta t_n} - \left(\tfrac{1}{\theta} - 1\right) \dot{\psi}_{l,n} \right) + \\
&\quad \psi^\tau_{l,n+1} \frac{\partial D_{il}(\boldsymbol{s}^\tau_{n+1})}{\partial s^\tau_{j,n+1}} - \\
&\quad \frac{\partial F_i(\boldsymbol{s}^\tau_{n+1})}{\partial s^\tau_{j,n+1}}
\end{aligned}
\tag{J.9}
$$

where the matrices $\boldsymbol{O}$, $\boldsymbol{B}$, $\boldsymbol{D}$ and the RHS-vector $\boldsymbol{F}$ are defined in (10.61). The partial Jacobians in (J.9) are obtained as follows

$$
J^{s1}_{ij} = \left(\frac{O_{ij}(\boldsymbol{s}^\tau_{n+1})}{\theta \Delta t_n} + D_{ij}(\boldsymbol{\psi}^\tau_{n+1}) \right) C^{-1^\tau}_{j,n+1} \qquad \text{no summation over } j
\tag{J.10}
$$

$$
J^{s2}_{ij} = \frac{B_{ij}}{\theta \Delta t_n}
\tag{J.11}
$$

$$
J^{s3}_{ij} = \sum_e \int_{\Omega^e} N_i \, S^e_o \delta_{ij} \left(\frac{\psi^\tau_{j,n+1} - \psi_{j,n}}{\theta \Delta t_n} - \left(\tfrac{1}{\theta} - 1\right) \dot{\psi}_{j,n} \right) d\Omega^e \qquad \text{no summation over } i \text{ and } j
\tag{J.12}
$$

$$
J^{s4}_{ij} = \sum_e \int_{\Omega^e} \nabla N_i \cdot \left(\boldsymbol{K}^e f^e_\mu N_j G^\tau_{j,n+1} C^{-1^\tau}_{j,n+1} \cdot \nabla N_l \psi^\tau_{l,n+1} \right) d\Omega^e \qquad \text{no summation over } j
\tag{J.13}
$$

$$
\begin{aligned}
J^{s5}_{ij} = -\sum_e \Bigg(&\int_{\Omega^e} \nabla N_i \cdot \left(\boldsymbol{K}^e f^e_\mu N_j (1 + \chi^e) G^\tau_{j,n+1} C^{-1^\tau}_{j,n+1} \cdot \boldsymbol{e} \right) d\Omega^e + \\
&\int_{\Gamma^{\nabla e}_N} N_i N_j G^\tau_{j,n+1} C^{-1^\tau}_{j,n+1} q^{\nabla e}_h \, d\Gamma^e \Bigg) \\
&\text{no summation over } j
\end{aligned}
\tag{J.14}
$$

with the inverse moisture capacity

$$C_{j,n+1}^{-1^{\tau}} = \frac{\partial \psi(s_{j,n+1}^{\tau})}{\partial s_{j,n+1}^{\tau}} = \frac{1}{C_{j,n+1}^{\tau}} \tag{J.15}$$

which can be evaluated either analytically from the parametric models summarized in Appendix D or numerically by using chord slope approximations given in Sect. J.3. Notice, it is necessary to use the pressure head ψ instead of the hydraulic head h to evaluate the moisture capacity functions $C_{j,n+1}^{\tau}$ and $C_{j,n+1}^{-1^{\tau}}$. Actually, $C_{j,n+1}^{\tau}$ can also be expressed by h since $\partial s/\partial\psi = \partial s/\partial h$, but the inverse moisture capacity $C_{j,n+1}^{-1^{\tau}}$ is not simply invertible for h because $\partial\psi/\partial s = \partial h/\partial s - \partial z/\partial s$.

It is important to note that in the derivation of the residual $\boldsymbol{R}$ with respect to the saturation $\boldsymbol{s}$ no assumptions are implied for spatial derivations of inherent parameters. Thus, (J.9) is also valid for inhomogeneous porous media. This is an essential difference to transformations for developing the common $s-$form of the Richards' equation as discussed in Sect. 10.3 or the Kirchhoff transformation as introduced in Sect. 10.4.

J.3 Chord Slope Approximations of Saturation Derivatives

In contrast to analytical derivatives in form of the moisture capacity $\boldsymbol{C}_{n+1}^{\tau}$ (J.7) and its inverse $\boldsymbol{C}_{n+1}^{-1^{\tau}}$ (J.15) chord slope approximations can be useful and effective. Within the GLS predictor-corrector one-step Newton scheme the derivative terms are evaluated by using the predicted solutions $\boldsymbol{\psi}_{n+1}^{p}$ and $\boldsymbol{s}_{n+1}^{p}$ for the current time plane $n+1$. For instance, a simple first-order accurate finite difference approximation of $\boldsymbol{C}_{n+1}^{\tau}$ would lead to

$$\boldsymbol{C}_{i,n+1}^{\tau} = \frac{s_{i,n+1}^{\tau} - s_{i,n}}{\psi_{i,n+1}^{\tau} - \psi_{i,n}} \tag{J.16}$$

For the GLS predictor-corrector one-step Newton technique only one iteration per time step is employed. Thus, the iterates indicated by the superscript τ can be replaced by the predictors denoted by the superscript P. This yields

$$\boldsymbol{C}_{i,n+1}^{p} = \frac{s_{i,n+1}^{p} - s_{i,n}}{\psi_{i,n+1}^{p} - \psi_{i,n}} \tag{J.17}$$

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

$$\boldsymbol{C}_{i,n+1}^{p} = \frac{\dot{s}_{i,n}}{\dot{\psi}_{i,n}} \tag{J.18}$$

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 [141]:

$$\boldsymbol{C}_{i,n+1}^{p} = \frac{\Delta t_{n-1}^{2}(s_{i,n+1}^{p} - s_{i,n}) + \Delta t_{n}^{2}(s_{i,n} - s_{i,n-1})}{\Delta t_{n-1}^{2}(\psi_{i,n+1}^{p} - \psi_{i,n}) + \Delta t_{n}^{2}(\psi_{i,n} - \psi_{i,n-1})} \tag{J.19}$$

The chord slope approximations for the inverse moisture capacity $\boldsymbol{C}_{i,n+1}^{-1^{p}}$ yield equivalent expressions.

Note here that limitations exist for the chord slope approximations if the denominator of (J.18) and (J.19) 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.

Appendix K
Integral Functions of the Frolkovič-Knabner Algorithm (FKA)

K.1 Transformations in Local Coordinates

The FKA computations are performed on generalized (local) coordinates $\boldsymbol{\eta}$ (8.68). The mapping from the local coordinates $\boldsymbol{\eta}$ to the global ones $\boldsymbol{x}$ is given by (8.114) and requires the transformation Jacobian $\boldsymbol{J}^e$, (8.115)–(8.117), for an element e. Using this transformation we have

$$\nabla_{(\xi,\eta,\zeta)} = \begin{pmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \\ \frac{\partial}{\partial \zeta} \end{pmatrix} = \boldsymbol{J}^e \cdot \nabla, \qquad \nabla = (\boldsymbol{J}^e)^{-1} \cdot \nabla_{(\xi,\eta,\zeta)} \tag{K.1}$$

for the derivatives and

$$\boldsymbol{e}_{(\xi,\eta,\zeta)} = \begin{pmatrix} e_\xi \\ e_\eta \\ e_\zeta \end{pmatrix} = \boldsymbol{J}^e \cdot \boldsymbol{e}, \qquad \boldsymbol{e} = (\boldsymbol{J}^e)^{-1} \cdot \boldsymbol{e}_{(\xi,\eta,\zeta)} \tag{K.2}$$

for the gravitational unit vector (3.261). Using these relationships the equivalent formulation of the Darcy velocity $\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot (\nabla h + \chi \boldsymbol{e})$ in local coordinates is given by

$$\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot \Big((\boldsymbol{J}^e)^{-1} \cdot (\nabla_{(\xi,\eta,\zeta)} h + \chi \boldsymbol{J}^e \cdot \boldsymbol{e}) \Big) \tag{K.3}$$

or

$$\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot \Big((\boldsymbol{J}^e)^{-1} \cdot (\nabla_{(\xi,\eta,\zeta)} h + \chi \boldsymbol{e}_{(\xi,\eta,\zeta)}) \Big) \tag{K.4}$$

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

K.2 FKA Formulation

Introducing the following integral functions [177, 312] for each element e

$$\begin{aligned} H_\xi^e &= H_\xi^e(\xi,\eta,\zeta) = \int_0^\xi \chi(\theta,\eta,\zeta) e_\xi(\theta,\eta,\zeta) d\theta \\ H_\eta^e &= H_\eta^e(\xi,\eta,\zeta) = \int_0^\eta \chi(\xi,\theta,\zeta) e_\eta(\xi,\theta,\zeta) d\theta \\ H_\zeta^e &= H_\zeta^e(\xi,\eta,\zeta) = \int_0^\zeta \chi(\xi,\eta,\theta) e_\zeta(\xi,\eta,\theta) d\theta \end{aligned} \tag{K.5}$$

and since

$$\begin{pmatrix} \frac{\partial H_\xi^e}{\partial \xi} \\ \frac{\partial H_\eta^e}{\partial \eta} \\ \frac{\partial H_\zeta^e}{\partial \zeta} \end{pmatrix} = \chi \boldsymbol{e}_{(\xi,\eta,\zeta)} \tag{K.6}$$

we can write the Darcy velocity (K.4) in an equivalent form on element level

$$\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot \left((\boldsymbol{J}^e)^{-1} \cdot \begin{pmatrix} \frac{\partial}{\partial \xi}(h + H_\xi^e) \\ \frac{\partial}{\partial \eta}(h + H_\eta^e) \\ \frac{\partial}{\partial \zeta}(h + H_\zeta^e) \end{pmatrix} \right) \tag{K.7}$$

The integral functions H_ξ^e, H_η^e, H_ζ^e allow us to obtain the same spatial variability for both the $h-$term and the $\chi-$term.

The consistency for (K.7) in the definition of (11.60) can be proved. Assuming the gravity acts in the $z-$direction, i.e., $\chi(x,y,z) = \chi(x_0,y_0,z)$, we can write

$$H_\xi^e = \int_0^\xi \chi(x_0,y_0,z(\theta,\eta,\zeta)) e_\xi d\theta = e_z \int_{z_0}^{z(\xi,\eta,\zeta)} \chi(x_0,y_0,\theta) d\theta \tag{K.8}$$

where $(x_0,y_0,z_0) = (x(0,0,0), y(0,0,0), z(0,0,0))$ and similarly for H_η^e and H_ζ^e.

In the FEM the functions h, H_ξ^e, H_η^e, H_ζ^e are interpolated by their nodal basis functions:

$$h = \sum_{J=1}^{N_{\text{BN}}} N_J^e(\xi,\eta,\zeta)\, h_J^e$$

$$H_\xi^e = \sum_{J=1}^{N_{\text{BN}}} N_J^e(\xi,\eta,\zeta)\, H_{\xi J}^e$$

$$H^e_\eta = \sum_{J=1}^{N_{\mathrm{BN}}} N^e_J(\xi,\eta,\zeta)\, H^e_{\eta J}$$

$$H^e_\zeta = \sum_{J=1}^{N_{\mathrm{BN}}} N^e_J(\xi,\eta,\zeta)\, H^e_{\zeta J} \tag{K.9}$$

and we obtain the velocity (K.7) in the discretized formulation

$$\boldsymbol{q} = -k_r \boldsymbol{K} f_\mu \cdot \Big((\boldsymbol{J}^e)^{-1} \cdot \sum_J^{N_{\mathrm{BN}}} \begin{pmatrix} (h^e_J + H^e_{\xi J}) \frac{\partial}{\partial \xi} N^e_J(\xi,\eta,\zeta) \\ (h^e_J + H^e_{\eta J}) \frac{\partial}{\partial \eta} N^e_J(\xi,\eta,\zeta) \\ (h^e_J + H^e_{\zeta J}) \frac{\partial}{\partial \zeta} N^e_J(\xi,\eta,\zeta) \end{pmatrix} \Big) \tag{K.10}$$

which represents a fully consistent approximation of the Darcy velocities. We solve (K.10) for given hydraulic heads h and the values of the $H^e_\xi, H^e_\eta, H^e_\zeta$–functions at the nodes J. The nodal quantities $H^e_{\xi J}, H^e_{\eta J}, H^e_{\zeta J}$ are dependent on the finite element types and will be evaluated next for linear elements in two and three dimensions. In doing this, the buoyancy coefficient χ in the gravity term is interpolated according to

$$\chi = \sum_{J=1}^{N_{\mathrm{BN}}} N^e_J(\xi,\eta,\zeta)\, \chi^e_J \tag{K.11}$$

where χ^e_J are the buoyancy coefficient values at node J. Using χ from (11.2) the finite element expansion of the buoyancy coefficient reads

$$\chi = \sum_k \tfrac{\alpha_k}{C_{ks}-C_{k0}} \Big(\sum_J N^e_J C^e_{kJ} - C_{k0}\Big) - \beta(T^e)\Big(\sum_J N^e_J T^e_J - T_0\Big) \tag{K.12}$$

or

$$\chi^e_J = \sum_k \tfrac{\alpha_k}{C_{ks}-C_{k0}} (C^e_{kJ} - C_{k0}) - \beta(T^e)(T^e_J - T_0) \tag{K.13}$$

in relation to the expansion (K.11)

K.3 The Nodal Quantities $H^e_{\xi J}$, $H^e_{\eta J}$, $H^e_{\zeta J}$ of the Integral Functions

K.3.1 2D Linear Triangular Element

The linear triangle is described in Sect. G.2 of Appendix G, where its shape functions and its local derivatives are given in Table G.2a. Its Jacobian $\boldsymbol{J}^e$ (H.14) appears independent of the local coordinates (ξ,η) and the gravitational unit vector $\boldsymbol{e}_{(\xi,\eta)}$ in the local coordinates (K.2) is also constant. Accordingly, we can write

$$
\begin{aligned}
H_\xi^e &= \int_0^\xi \chi(\theta,\eta) e_\xi(\theta,\eta) d\theta = e_\xi \int_0^\xi \chi(\theta,\eta) d\theta \\
&= e_\xi \int_0^\xi \Big(\sum_{J=1}^{3} N_J^e(\theta,\eta) \chi_J^e \Big) d\theta = \\
&= e_\xi \int_0^\xi \big[(1-\theta-\eta)\chi_1^e + \theta\chi_2^e + \eta\chi_3^e \big] d\theta \\
&= e_\xi \big[(\xi - \tfrac{\xi^2}{2} - \xi\eta)\chi_1^e + \tfrac{\xi^2}{2}\chi_2^e + \xi\eta\chi_3^e \big]
\end{aligned}
\tag{K.14}
$$

and similarly for H_η^e. From the integrals we find the nodal values for H_ξ^e and H_η^e as

$$
\begin{aligned}
H_\xi^e(0,0) &= H_{\xi 1}^e = 0 \\
H_\xi^e(1,0) &= H_{\xi 2}^e = \tfrac{1}{2} e_\xi (\chi_1^e + \chi_2^e) \\
H_\xi^e(0,1) &= H_{\xi 3}^e = 0
\end{aligned}
\tag{K.15}
$$

$$
\begin{aligned}
H_\eta^e(0,0) &= H_{\eta 1}^e = 0 \\
H_\eta^e(1,0) &= H_{\eta 2}^e = 0 \\
H_\eta^e(0,1) &= H_{\eta 3}^e = \tfrac{1}{2} e_\eta (\chi_1^e + \chi_3^e)
\end{aligned}
\tag{K.16}
$$

Now we can express the gravity term (K.6) in local coordinates as

$$
\begin{aligned}
\chi \boldsymbol{e}_{(\xi,\eta)} &= \begin{pmatrix} \frac{\partial H_\xi^e}{\partial \xi} \\ \frac{\partial H_\eta^e}{\partial \eta} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\partial N_1^e}{\partial \xi} H_{\xi 1}^e + \frac{\partial N_2^e}{\partial \xi} H_{\xi 2}^e + \frac{\partial N_3^e}{\partial \xi} H_{\xi 3}^e \\ \frac{\partial N_1^e}{\partial \eta} H_{\eta 1}^e + \frac{\partial N_2^e}{\partial \eta} H_{\eta 2}^e + \frac{\partial N_3^e}{\partial \eta} H_{\eta 3}^e \end{pmatrix} \\
&= \frac{1}{2} \begin{pmatrix} e_\xi (\chi_1^e + \chi_2^e) \\ e_\eta (\chi_1^e + \chi_3^e) \end{pmatrix}
\end{aligned}
\tag{K.17}
$$

representing a consistent approximation in which the density is appropriately averaged in the gravitational direction.

K.3.2 2D Linear Quadrilateral Element

The linear quadrilateral is described in Sect. G.2 of Appendix G, where its shape functions and its local derivatives are given in Table G.2b. While for this element the Jacobian $\boldsymbol{J}^e$ (H.49) is in general space-dependent, the gravitational unit vector $\boldsymbol{e}_{(\xi,\eta)}$ in local coordinates (K.2) takes the special form

$$
\begin{pmatrix} e_\xi \\ e_\eta \end{pmatrix} = \begin{pmatrix} e_\xi(\eta) \\ e_\eta(\xi) \end{pmatrix}
\tag{K.18}
$$

Similarly to the above triangular element, we can compute the integral functions H^e_ξ, H^e_η at the corner nodes J for the linear quadrilateral element as

$$\begin{aligned}
H^e_\xi(-1,-1) &= H^e_{\xi 1} = -\tfrac{1}{4} e_\xi(-1)(3\chi^e_1 + \chi^e_2) \\
H^e_\xi(1,-1) &= H^e_{\xi 2} = \tfrac{1}{4} e_\xi(-1)(\chi^e_1 + 3\chi^e_2) \\
H^e_\xi(1,1) &= H^e_{\xi 3} = \tfrac{1}{4} e_\xi(1)(3\chi^e_3 + \chi^e_4) \\
H^e_\xi(-1,1) &= H^e_{\xi 4} = -\tfrac{1}{4} e_\xi(1)(\chi^e_3 + 3\chi^e_4)
\end{aligned} \tag{K.19}$$

$$\begin{aligned}
H^e_\eta(-1,-1) &= H^e_{\eta 1} = -\tfrac{1}{4} e_\eta(-1)(3\chi^e_1 + \chi^e_4) \\
H^e_\eta(1,-1) &= H^e_{\eta 2} = -\tfrac{1}{4} e_\eta(1)(3\chi^e_2 + \chi^e_3) \\
H^e_\eta(1,1) &= H^e_{\eta 3} = \tfrac{1}{4} e_\eta(1)(\chi^e_2 + 3\chi^e_3) \\
H^e_\eta(-1,1) &= H^e_{\eta 4} = \tfrac{1}{4} e_\eta(-1)(\chi^e_1 + 3\chi^e_4)
\end{aligned} \tag{K.20}$$

The gravity term (K.6) written in local coordinates yields

$$\begin{aligned}
\chi \boldsymbol{e}_{(\xi,\eta)} &= \begin{pmatrix} \frac{\partial H^e_\xi}{\partial \xi} \\ \frac{\partial H^e_\eta}{\partial \eta} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\partial N^e_1}{\partial \xi} H^e_{\xi 1} + \frac{\partial N^e_2}{\partial \xi} H^e_{\xi 2} + \frac{\partial N^e_3}{\partial \xi} H^e_{\xi 3} + \frac{\partial N^e_4}{\partial \xi} H^e_{\xi 4} \\ \frac{\partial N^e_1}{\partial \eta} H^e_{\eta 1} + \frac{\partial N^e_2}{\partial \eta} H^e_{\eta 2} + \frac{\partial N^e_3}{\partial \eta} H^e_{\eta 3} + \frac{\partial N^e_4}{\partial \eta} H^e_{\eta 4} \end{pmatrix} \\
&= \tfrac{1}{4} \begin{pmatrix} e_\xi(-1)(1-\eta)(\chi^e_1 + \chi^e_2) + e_\xi(1)(1+\eta)(\chi^e_3 + \chi^e_4) \\ e_\eta(-1)(1-\xi)(\chi^e_1 + \chi^e_4) + e_\eta(1)(1+\xi)(\chi^e_2 + \chi^e_3) \end{pmatrix}
\end{aligned} \tag{K.21}$$

For the linear quadrilateral element the consistent approximation (K.21) can be recognized as the consistent formulation previously introduced by Voss [550], where the gravity term is averaged in a directional manner, so for instance

$$\chi \boldsymbol{e}_{(\xi=1,\eta=1)} = \tfrac{1}{2} \begin{pmatrix} e_\xi(1)(\chi^e_3 + \chi^e_4) \\ e_\eta(1)(\chi^e_2 + \chi^e_3) \end{pmatrix} \tag{K.22}$$

K.3.3 3D Linear Pentahedral (Triangular Prismatic) Element

The linear pentahedral (triangular prismatic) element is described in Sect. G.3 of Appendix G, where its shape functions and its local derivatives are given in Table G.3b. Since the Jacobian $\boldsymbol{J}^e$ (H.67) is space-dependent, the gravitational unit vector $\boldsymbol{e}_{(\xi,\eta,\zeta)}$ in local coordinates (K.2) is

$$\begin{pmatrix} e_\xi \\ e_\eta \\ e_\zeta \end{pmatrix} = \begin{pmatrix} e_\xi(\zeta) \\ e_\eta(\zeta) \\ e_\zeta(\xi,\eta) \end{pmatrix} \tag{K.23}$$

The integral functions H_ξ^e, H_η^e, H_ζ^e at the corner nodes J for the linear pentahedral element are then

$$\begin{aligned}
H_\xi^e(0,0,1) &= H_{\xi 1}^e = 0\\
H_\xi^e(1,0,1) &= H_{\xi 2}^e = \tfrac{1}{2} e_\xi(1)(\chi_1^e + \chi_2^e)\\
H_\xi^e(0,1,1) &= H_{\xi 3}^e = 0\\
H_\xi^e(0,0,-1) &= H_{\xi 4}^e = 0\\
H_\xi^e(1,0,-1) &= H_{\xi 5}^e = \tfrac{1}{2} e_\xi(-1)(\chi_5^e + \chi_6^e)\\
H_\xi^e(0,1,-1) &= H_{\xi 6}^e = 0
\end{aligned} \tag{K.24}$$

$$\begin{aligned}
H_\eta^e(0,0,1) &= H_{\eta 1}^e = 0\\
H_\eta^e(1,0,1) &= H_{\eta 2}^e = 0\\
H_\eta^e(0,1,1) &= H_{\eta 3}^e = \tfrac{1}{2} e_\eta(1)(\chi_1^e + \chi_3^e)\\
H_\eta^e(0,0,-1) &= H_{\eta 4}^e = 0\\
H_\eta^e(1,0,-1) &= H_{\eta 5}^e = 0\\
H_\eta^e(0,1,-1) &= H_{\eta 6}^e = \tfrac{1}{2} e_\eta(-1)(\chi_4^e + \chi_6^e)
\end{aligned} \tag{K.25}$$

$$\begin{aligned}
H_\zeta^e(0,0,1) &= H_{\zeta 1}^e = \tfrac{1}{4} e_\zeta(0,0)(3\chi_1^e + \chi_4^e)\\
H_\zeta^e(1,0,1) &= H_{\zeta 2}^e = \tfrac{1}{4} e_\zeta(1,0)(3\chi_2^e + \chi_5^e)\\
H_\zeta^e(0,1,1) &= H_{\zeta 3}^e = \tfrac{1}{4} e_\zeta(0,1)(3\chi_3^e + \chi_6^e)\\
H_\zeta^e(0,0,-1) &= H_{\zeta 4}^e = -\tfrac{1}{4} e_\zeta(0,0)(\chi_1^e + 3\chi_4^e)\\
H_\zeta^e(1,0,-1) &= H_{\zeta 5}^e = -\tfrac{1}{4} e_\zeta(1,0)(\chi_2^e + 3\chi_5^e)\\
H_\zeta^e(0,1,-1) &= H_{\zeta 6}^e = -\tfrac{1}{4} e_\zeta(0,1)(\chi_3^e + 3\chi_6^e)
\end{aligned} \tag{K.26}$$

The gravity term (K.6) written in local coordinates gives

$$\begin{aligned}
\chi \boldsymbol{e}_{(\xi,\eta)} &= \begin{pmatrix} \frac{\partial H_\xi^e}{\partial \xi} \\ \frac{\partial H_\eta^e}{\partial \eta} \\ \frac{\partial H_\zeta^e}{\partial \zeta} \end{pmatrix}\\
&= \begin{pmatrix}
\frac{\partial N_1^e}{\partial \xi} H_{\xi 1}^e + \frac{\partial N_2^e}{\partial \xi} H_{\xi 2}^e + \frac{\partial N_3^e}{\partial \xi} H_{\xi 3}^e + \frac{\partial N_4^e}{\partial \xi} H_{\xi 4}^e + \frac{\partial N_5^e}{\partial \xi} H_{\xi 5}^e + \frac{\partial N_6^e}{\partial \xi} H_{\xi 6}^e \\
\frac{\partial N_1^e}{\partial \eta} H_{\eta 1}^e + \frac{\partial N_2^e}{\partial \eta} H_{\eta 2}^e + \frac{\partial N_3^e}{\partial \eta} H_{\eta 3}^e + \frac{\partial N_4^e}{\partial \eta} H_{\eta 4}^e + \frac{\partial N_5^e}{\partial \eta} H_{\eta 5}^e + \frac{\partial N_6^e}{\partial \eta} H_{\eta 6}^e \\
\frac{\partial N_1^e}{\partial \zeta} H_{\zeta 1}^e + \frac{\partial N_2^e}{\partial \zeta} H_{\zeta 2}^e + \frac{\partial N_3^e}{\partial \zeta} H_{\zeta 3}^e + \frac{\partial N_4^e}{\partial \zeta} H_{\zeta 4}^e + \frac{\partial N_5^e}{\partial \zeta} H_{\zeta 5}^e + \frac{\partial N_6^e}{\partial \zeta} H_{\zeta 6}^e
\end{pmatrix}\\
&= \frac{1}{4} \begin{pmatrix}
e_\xi(1)(1+\zeta)(\chi_1^e + \chi_2^e) + e_\xi(-1)(1-\zeta)(\chi_5^e + \chi_6^e) \\
e_\eta(1)(1+\zeta)(\chi_1^e + \chi_3^e) + e_\eta(-1)(1-\zeta)(\chi_4^e + \chi_6^e) \\
2[e_\zeta(0,0)(1-\xi-\eta)(\chi_1^e + \chi_4^e) + e_\zeta(1,0)\xi(\chi_2^e + \chi_5^e) + e_\zeta(0,1)\eta(\chi_3^e + \chi_6^e)]
\end{pmatrix}
\end{aligned} \tag{K.27}$$

K.3.4 3D Linear Hexahedral (Brick) Element

The linear hexahedral (brick) element is described in Sect. G.3 of Appendix G, where its shape functions and its local derivatives are given in Table G.3c. The Jacobian $\boldsymbol{J}^e$ (H.57) is space-dependent and the gravitational unit vector $\boldsymbol{e}_{(\xi,\eta,\zeta)}$ in local coordinates (K.2) reads

$$\begin{pmatrix} e_\xi \\ e_\eta \\ e_\zeta \end{pmatrix} = \begin{pmatrix} e_\xi(\eta,\zeta) \\ e_\eta(\xi,\zeta) \\ e_\zeta(\xi,\eta) \end{pmatrix} \tag{K.28}$$

The integral functions H^e_ξ, H^e_η, H^e_ζ at the corner nodes J for the linear hexahedral element can be derived as

$$\begin{aligned}
H^e_\xi(-1,-1,1) &= H^e_{\xi 1} = -\tfrac{1}{4}e_\xi(-1,1)(3\chi^e_1 + \chi^e_2) \\
H^e_\xi(1,-1,1) &= H^e_{\xi 2} = \tfrac{1}{4}e_\xi(-1,1)(\chi^e_1 + 3\chi^e_2) \\
H^e_\xi(1,1,1) &= H^e_{\xi 3} = \tfrac{1}{4}e_\xi(1,1)(3\chi^e_3 + \chi^e_4) \\
H^e_\xi(-1,1,1) &= H^e_{\xi 4} = -\tfrac{1}{4}e_\xi(1,1)(\chi^e_3 + 3\chi^e_4) \\
H^e_\xi(-1,-1,-1) &= H^e_{\xi 5} = -\tfrac{1}{4}e_\xi(-1,-1)(3\chi^e_5 + \chi^e_6) \\
H^e_\xi(1,-1,-1) &= H^e_{\xi 6} = \tfrac{1}{4}e_\xi(-1,-1)(\chi^e_5 + 3\chi^e_6) \\
H^e_\xi(1,1,-1) &= H^e_{\xi 7} = \tfrac{1}{4}e_\xi(1,-1)(3\chi^e_7 + \chi^e_8) \\
H^e_\xi(-1,1,-1) &= H^e_{\xi 8} = -\tfrac{1}{4}e_\xi(1,-1)(\chi^e_7 + 3\chi^e_8)
\end{aligned} \tag{K.29}$$

$$\begin{aligned}
H^e_\eta(-1,-1,1) &= H^e_{\eta 1} = -\tfrac{1}{4}e_\eta(-1,1)(3\chi^e_1 + \chi^e_4) \\
H^e_\eta(1,-1,1) &= H^e_{\eta 2} = -\tfrac{1}{4}e_\eta(1,1)(3\chi^e_2 + \chi^e_3) \\
H^e_\eta(1,1,1) &= H^e_{\eta 3} = \tfrac{1}{4}e_\eta(1,1)(\chi^e_2 + 3\chi^e_3) \\
H^e_\eta(-1,1,1) &= H^e_{\eta 4} = \tfrac{1}{4}e_\eta(-1,1)(\chi^e_1 + 3\chi^e_4) \\
H^e_\eta(-1,-1,-1) &= H^e_{\eta 5} = -\tfrac{1}{4}e_\eta(-1,-1)(3\chi^e_5 + \chi^e_8) \\
H^e_\eta(1,-1,-1) &= H^e_{\eta 6} = -\tfrac{1}{4}e_\eta(1,-1)(3\chi^e_6 + \chi^e_7) \\
H^e_\eta(1,1,-1) &= H^e_{\eta 7} = \tfrac{1}{4}e_\eta(1,-1)(\chi^e_6 + 3\chi^e_7) \\
H^e_\eta(-1,1,-1) &= H^e_{\eta 8} = \tfrac{1}{4}e_\eta(-1,-1)(\chi^e_5 + 3\chi^e_8)
\end{aligned} \tag{K.30}$$

$$\begin{aligned}
H^e_\zeta(-1,-1,1) &= H^e_{\zeta 1} = \tfrac{1}{4}e_\zeta(-1,-1)(3\chi^e_1 + \chi^e_5) \\
H^e_\zeta(1,-1,1) &= H^e_{\zeta 2} = \tfrac{1}{4}e_\zeta(1,-1)(3\chi^e_2 + \chi^e_6) \\
H^e_\zeta(1,1,1) &= H^e_{\zeta 3} = \tfrac{1}{4}e_\zeta(1,1)(3\chi^e_3 + \chi^e_7) \\
H^e_\zeta(-1,1,1) &= H^e_{\zeta 4} = \tfrac{1}{4}e_\zeta(-1,1)(3\chi^e_4 + \chi^e_8) \\
H^e_\zeta(-1,-1,-1) &= H^e_{\zeta 5} = -\tfrac{1}{4}e_\zeta(-1,-1)(\chi^e_1 + 3\chi^e_5) \\
H^e_\zeta(1,-1,-1) &= H^e_{\zeta 6} = -\tfrac{1}{4}e_\zeta(1,-1)(\chi^e_2 + 3\chi^e_6) \\
H^e_\zeta(1,1,-1) &= H^e_{\zeta 7} = -\tfrac{1}{4}e_\zeta(1,1)(\chi^e_3 + 3\chi^e_7) \\
H^e_\zeta(-1,1,-1) &= H^e_{\zeta 8} = -\tfrac{1}{4}e_\zeta(-1,1)(\chi^e_4 + 3\chi^e_8)
\end{aligned} \tag{K.31}$$

For the hexahedral (brick) element the consistent formulation of the gravity term in form of the integral functions, (K.29)–(K.31), is equivalent to the formulation given by Leijnse [336]. This should be exemplified for the χe_ξ−component of the gravity term:

$$\begin{aligned}\chi e_\xi = \textstyle\sum_{J=1}^{8} \frac{\partial N_J^e}{\partial \xi} H_{\xi J}^e = \tfrac{1}{8}\big[&e_\xi(-1,1)(1-\eta)(1+\zeta)(\chi_1^e+\chi_2^e)+\\ &e_\xi(1,1)(1+\eta)(1+\zeta)(\chi_3^e+\chi_4^e)+\\ &e_\xi(-1,-1)(1-\eta)(1-\zeta)(\chi_5^e+\chi_6^e)+\\ &e_\xi(1,-1)(1+\eta)(1-\zeta)(\chi_7^e+\chi_8^e)\big]\end{aligned} \tag{K.32}$$

K.3.5 3D Linear Pyramidal Element

The linear pyramidal element is described in Sect. G.3 of Appendix G, where its shape functions and its local derivatives are given in Table G.3d. Since the Jacobian $\boldsymbol{J}^e$ (H.78) is space-dependent, the gravitational unit vector $\boldsymbol{e}_{(\xi,\eta,\zeta)}$ in local coordinates (K.2) is

$$\begin{pmatrix} e_\xi \\ e_\eta \\ e_\zeta \end{pmatrix} = \begin{pmatrix} e_\xi(\eta,\zeta) \\ e_\eta(\xi,\zeta) \\ e_\zeta(\xi,\eta,\zeta) \end{pmatrix} \tag{K.33}$$

The integral functions H_ξ^e, H_η^e, H_ζ^e at the corner nodes J for the linear pyramidal element are then

$$\begin{aligned} H_\xi^e(-1,-1,0) &= H_{\xi 1}^e = -\tfrac{1}{4}e_\xi(-1,0)(3\chi_1^e+\chi_2^e)\\ H_\xi^e(1,-1,0) &= H_{\xi 2}^e = \tfrac{1}{4}e_\xi(-1,0)(\chi_1^e+3\chi_2^e)\\ H_\xi^e(1,1,0) &= H_{\xi 3}^e = \tfrac{1}{4}e_\xi(1,0)(3\chi_3^e+\chi_4^e)\\ H_\xi^e(-1,1,0) &= H_{\xi 4}^e = -\tfrac{1}{4}e_\xi(1,0)(\chi_3^e+3\chi_4^e)\\ H_\xi^e(0,0,1) &= H_{\xi 5}^e = 0 \end{aligned} \tag{K.34}$$

$$\begin{aligned} H_\eta^e(-1,-1,0) &= H_{\eta 1}^e = -\tfrac{1}{4}e_\eta(-1,0)(3\chi_1^e+\chi_4^e)\\ H_\eta^e(1,-1,0) &= H_{\eta 2}^e = -\tfrac{1}{4}e_\eta(1,0)(3\chi_2^e+\chi_3^e)\\ H_\eta^e(1,1,0) &= H_{\eta 3}^e = \tfrac{1}{4}e_\eta(1,0)(\chi_2^e+3\chi_3^e)\\ H_\eta^e(-1,1,0) &= H_{\eta 4}^e = \tfrac{1}{4}e_\eta(-1,0)(\chi_1^e+3\chi_4^e)\\ H_\eta^e(0,0,1) &= H_{\eta 5}^e = 0 \end{aligned} \tag{K.35}$$

$$\begin{aligned} H_\zeta^e(-1,-1,0) &= H_{\zeta 1}^e = \tfrac{1}{4}e_\zeta(-1,-1,0)(\chi_1^e-\chi_2^e+\chi_3^e-\chi_4^e)\\ H_\zeta^e(1,-1,0) &= H_{\zeta 2}^e = \tfrac{1}{4}e_\zeta(1,-1,0)(-\chi_1^e+\chi_2^e-\chi_3^e+\chi_4^e)\\ H_\zeta^e(1,1,0) &= H_{\zeta 3}^e = \tfrac{1}{4}e_\zeta(1,1,0)(\chi_1^e-\chi_2^e+\chi_3^e-\chi_4^e)\\ H_\zeta^e(-1,1,0) &= H_{\zeta 4}^e = \tfrac{1}{4}e_\zeta(-1,1,0)(-\chi_1^e+\chi_2^e-\chi_3^e+\chi_4^e)\\ H_\zeta^e(0,0,1) &= H_{\zeta 5}^e = \tfrac{1}{8}e_\zeta(0,0,1)(\chi_1^e+\chi_2^e+\chi_3^e+\chi_4^e+4\chi_5^e) \end{aligned} \tag{K.36}$$

Then, the gravity term (K.6) written in local coordinates reads

$$
\begin{aligned}
\chi \boldsymbol{e}_{(\xi,\eta)} &= \begin{pmatrix} \frac{\partial H^e_\xi}{\partial \xi} \\ \frac{\partial H^e_\eta}{\partial \eta} \\ \frac{\partial H^e_\zeta}{\partial \zeta} \end{pmatrix} \\
&= \begin{pmatrix}
\frac{\partial N^e_1}{\partial \xi} H^e_{\xi 1} + \frac{\partial N^e_2}{\partial \xi} H^e_{\xi 2} + \frac{\partial N^e_3}{\partial \xi} H^e_{\xi 3} + \frac{\partial N^e_4}{\partial \xi} H^e_{\xi 4} + \frac{\partial N^e_5}{\partial \xi} H^e_{\xi 5} \\
\frac{\partial N^e_1}{\partial \eta} H^e_{\eta 1} + \frac{\partial N^e_2}{\partial \eta} H^e_{\eta 2} + \frac{\partial N^e_3}{\partial \eta} H^e_{\eta 3} + \frac{\partial N^e_4}{\partial \eta} H^e_{\eta 4} + \frac{\partial N^e_5}{\partial \eta} H^e_{\eta 5} \\
\frac{\partial N^e_1}{\partial \zeta} H^e_{\zeta 1} + \frac{\partial N^e_2}{\partial \zeta} H^e_{\zeta 2} + \frac{\partial N^e_3}{\partial \zeta} H^e_{\zeta 3} + \frac{\partial N^e_4}{\partial \zeta} H^e_{\zeta 4} + \frac{\partial N^e_5}{\partial \zeta} H^e_{\zeta 5}
\end{pmatrix} \\
&= \tfrac{1}{16} \begin{pmatrix}
4[e_\xi(-1,0)(1-\eta-\frac{\eta\zeta}{1-\zeta})(\chi^e_1+\chi^e_2) + e_\xi(1,0)(1+\eta+\frac{\eta\zeta}{1-\zeta})(\chi^e_3+\chi^e_4)] \\
4[e_\eta(-1,0)(1-\xi-\frac{\xi\zeta}{1-\zeta})(\chi^e_1+\chi^e_4) + e_\eta(1,0)(1+\xi+\frac{\xi\zeta}{1-\zeta})(\chi^e_2+\chi^e_3)] \\
\big\{[e_\zeta(-1,-1,0)+e_\zeta(1,1,0)]\big(1-\frac{\xi\eta}{(1-\zeta)^2}\big)(-\chi^e_1+\chi^e_2-\chi^e_3+\chi^e_4)+ \\
[e_\zeta(1,-1,0)+e_\zeta(-1,1,0)]\big(1+\frac{\xi\eta}{(1-\zeta)^2}\big)(\chi^e_1-\chi^e_2+\chi^e_3-\chi^e_4)+ \\
2e_\zeta(0,0,1)(\chi^e_1+\chi^e_2+\chi^e_3+\chi^e_4+4\chi^e_5)\big\}
\end{pmatrix}
\end{aligned}
\tag{K.37}
$$

Appendix L
Formulation of Hydraulic Head BC's for Variable-Density Problems

L.1 Problem Description

In formulating flow BC's the prescription of a hydraulic head (first kind Dirichlet-type or third kind Cauchy-type) BC at a given boundary portion is a common task (Sect. 6.3.1). However, in modeling variable-density problems such as saltwater intrusion or geothermal processes these hydraulic head BC's have to consider the specific definition of the hydraulic head h (3.260), viz.,

$$h = \frac{p}{\rho_0 g} + z \tag{L.1}$$

which must be appropriately related to a *reference fluid density* ρ_0 (cf. Sect. 3.8.6.1). A typical example is the saltwater intrusion from a sea into a coastal aquifer as schematized in Fig. L.1, where at the sea side the boundary is imposed by a given hydraulic head distribution $h(z)$. On the other hand, at the sea the hydraulic head can be measured in form of a piezometric head h_s which is related to the actual fluid density of saltwater ρ_s, while the reference fluid density ρ_0 typically refers to the *freshwater*.

L.2 Reference Potential from Measured Heads

A measurement of a piezometric head is normally related to the actual fluid density. It can be expressed for saltwater by

$$h_s = \frac{p}{\rho_s g} + z \tag{L.2}$$

where ρ_s is the fluid density at known salinity C_s and temperature T_s: $\rho_s = \rho(C_s, T_s)$. (Note that the salinity C_s concerns a single-species concentration, where for the sake of simplicity we can drop the species index k.) It is obvious that

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

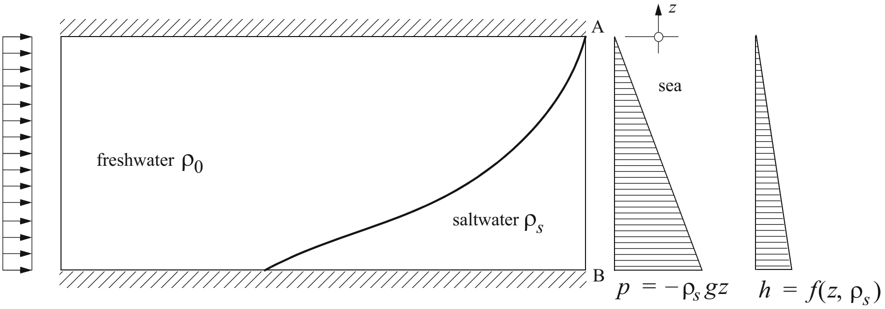


Fig. L.1 Saltwater intrusion in a coastal aquifer with related BC's

the piezometric head h_s cannot be directly used as a BC. Instead, it has to be transformed to the hydraulic head h (L.1), which is related to the reference fluid density ρ_0. This can be simply done under considering the following relationships:

Expanding (L.2) by ρ_0

$$h_s = \frac{p}{\rho_0 g}\frac{\rho_0}{\rho_s} + z \tag{L.3}$$

we obtain if introducing (L.1)

$$h_s = \frac{\rho_0}{\rho_s}h + \Big(1 - \frac{\rho_0}{\rho_s}\Big)z \tag{L.4}$$

and finally

$$h = \frac{\rho_s}{\rho_0}h_s - \Big(\frac{\rho_s - \rho_0}{\rho_0}\Big)z \tag{L.5}$$

For the fluid density relation we find from (3.265), (3.278) or (11.2)

$$\frac{\rho_s - \rho_0}{\rho_0} = \alpha - \beta(T_s)(T_s - T_0) \tag{L.6}$$

where α is the specific solutal expansion coefficient (density ratio) defined by (3.275) and $\beta(T_s)$ is the thermal expansion coefficient defined in (11.2). Inserting (L.6) into (L.5) it results

$$h = (1+\alpha)h_s - \alpha z + \beta(T_s)(T_s - T_0)(z - h_s) \tag{L.7}$$

The relation (L.7) is to be used to calculate the (freshwater) hydraulic heads h from (saltwater) piezometric heads h_s measured at a known saltwater density ρ_s (at known salinity C_s and temperature T_s). Under isothermal conditions (L.7) reduces to

$$h = (1+\alpha)h_s - \alpha z \tag{L.8}$$

L.3 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. L.2:

$$\rho = \rho_1 + (\rho_1 - \rho_2)\tfrac{z}{H}, \qquad -H \le z \le 0 \tag{L.9}$$

The fluid is hydrostatic for the vertical problem if

$$\begin{aligned} \frac{dp}{dz} &= -\rho g \\ p(z) &= -g \int_{z_1}^{z} \rho(\theta) d\theta \end{aligned} \tag{L.10}$$

which yields with (L.9)

$$p = p_1 - g\big(\rho_1 z + \tfrac{\rho_1 - \rho_2}{2H} z^2\big) \tag{L.11}$$

The hydraulic head h (L.1) related to the reference density ρ_0 is then

$$h = h_1 - \big(\tfrac{\rho_1 - \rho_0}{\rho_0}\big)z - \tfrac{1}{2H}\big(\tfrac{\rho_1 - \rho_2}{\rho_0}\big)z^2 \tag{L.12}$$

At boundaries where hydrostatic conditions can be imposed two cases are commonly of interest as illustrated in Fig. L.3:

1. A constant saltwater density in the depth and
2. A linear increase of density as typical in a transition zone.

From (L.12) we obtain with $\rho_1 = \rho_2 = \rho_s$ for the case of a constant saltwater density (case 1), written for the sake of simplicity under isothermal conditions

$$h = h_s - \alpha z \tag{L.13}$$

For the case 2 with $\rho_1 = \rho_0$, $\rho_2 = \rho_s$, we get from (L.12) for a linear saltwater density (case 2)

$$h = h_0 + \tfrac{\alpha}{2H} z^2 \tag{L.14}$$

The hydraulic head at the depth $z = -H$ is then $h = h_s + \alpha H$ for the constant density and $h = h_0 + \frac{\alpha}{2}H$ for the linear density relationship. Corresponding expressions can be derived for nonisothermal conditions if using (L.6), viz.,

$$h = h_s + [\alpha - \beta(T_s)(T_s - T_0)]H \tag{L.15}$$

Fig. L.2 Hydrostatic condition in a depth of H under a linear density gradient $\rho = \rho_1 + (\rho_1 - \rho_2)\frac{z}{H}$

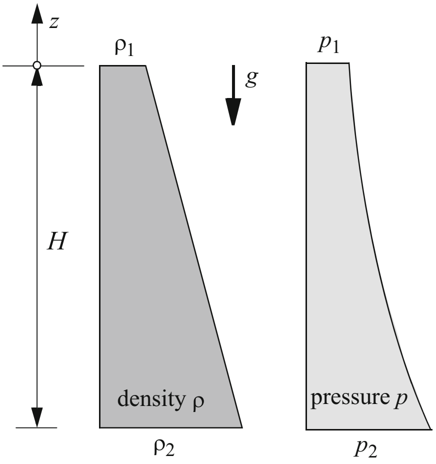
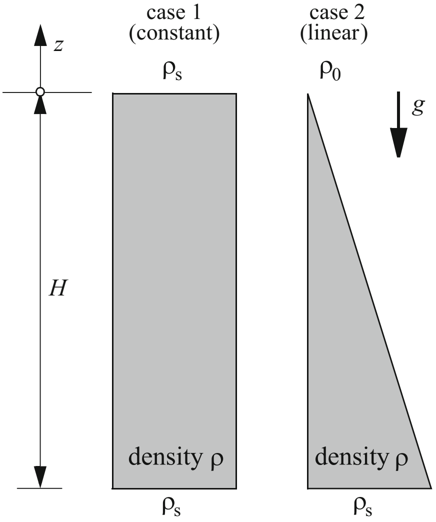


Fig. L.3 Two interesting density profiles for a hydrostatic BC



for the constant density relationship and

$$h = h_0 + \tfrac{1}{2}[\alpha - \beta(T_s)(T_s - T_0)]H \tag{L.16}$$

for the linear density relationship.

Based on these relations we are able to specify head conditions along boundaries with a given fluid density profile. Let us consider the example as shown in Fig. L.4 where a transition zone at a vertical boundary should be modeled for a saltwater intrusion process. The fluid density $\rho = \rho(z)$ varies linearly through the transition zone with a thickness of δ. At such a boundary a hydraulic head condition $h = h(z)$

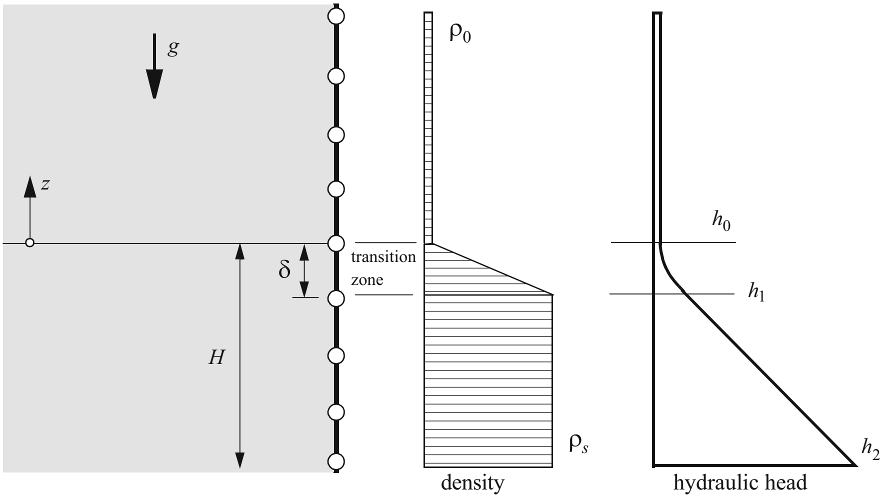


Fig. L.4 Boundary with a predefined saltwater-freshwater transition zone

has to be imposed. From (L.14) to (L.13) we obtain the following sample values for the hydraulic head profile under isothermal conditions as illustrated in Fig. L.4:

$$\begin{aligned} h_1 &= h_0 + \tfrac{\alpha}{2}\delta \\ h_2 &= h_1 + \alpha(H - \delta) = h_0 + \alpha\big(H - \tfrac{\delta}{2}\big) \end{aligned} \tag{L.17}$$

where h_0 represents the hydraulic head at the boundary which is related to the freshwater density ρ_0.

Appendix M
BHE Modeling: Numerical and Analytical Approaches

M.1 Types of BHE

In this section, different types of BHE with their individual pipe and grout components are described. They form highly slender cylindrical boreholes. The BHE systems are represented by 1D schematizations, where the pipe and grout components have a reduced spatial dimension. They imply that the variation of the temperature is along the vertical axis. The heat fluxes normal to the contact surfaces for the 1D pipe and grout components are modeled by heat transfer relations.

M.1.1 Double U-Shape Pipe (2U)

The double U-shape pipe (2U) exchanger is a cylindrical borehole consisting of two inner pipes forming a U-shape and filled with a grout material, cf. Fig. 13.1. Basically, the grout can be considered as a homogeneous impervious material and could be schematized by only one component so as proposed in [6–8]. However, to improve the approximation of the inner pipe-to-grout heat transfer we introduce a larger number of grout components in the extended TRCM, which correlates with the number of pipes of BHE [29]. This has some advantages: (1) A better accuracy results in modeling the transient behavior of U-shape pipe exchangers. (2) It allows a much higher flexibility in configuration of U-shape pipe systems, particularly, the U-shape pipes can be arranged crosswise or side by side. (3) Furthermore, the flow through double U-shape pipe configurations can be parallel or serial. In total, we schematize a 2U exchanger by eight components (Fig. M.1):

- Two pipes-in (denoted as $i1$ and $i2$)
- Two pipes-out (denoted as $o1$ and $o2$)
- Grout material, which is subdivided into four zones (denoted as $g1$, $g2$, $g3$, $g4$)

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

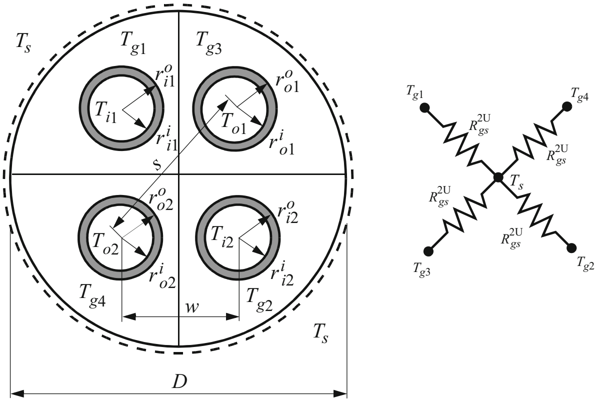


Fig. M.1 Inner pipe-grout heat flux resistance relationships of a 2U borehole consisting of four pipe components and four grout zones

The four pipe components $i1, i2, o1$ and $o2$ transfer heat across their cross-sectional areas and exchange fluxes across their surface areas. The radial heat transfer from the pipes is directed to the grout zones gi $(i = 1, \dots, 4)$. The grout zones gi $(i = 1, \dots, 4)$ exchange heat directly to the surrounding soil (the porous matrix with the filled fluid in the void space) denoted as s and to other contacted grout zones too. It can be seen that, as physically occurring, the heat coupling only occurs via the grout zones gi $(i = 1, \dots, 4)$, which work as intermediate media that transfer heat from one pipe to another and vice versa. Only the grout zones exchange heat with the surrounding soil s because there is no direct thermal contact between the pipes $i1, i2, o1$ and $o2$ with the soil s.

The 2U system involves several material and geometric parameters, which are either given by the manufacturer of the heating system or determined experimentally. These relations are used to express the overall thermal resistance between the 2U borehole and the soil. The usual practice is to lump the effects of the 2U components into effective heat transfer coefficients representing the reciprocal of the sum of the thermal resistances between the different components (cf. Appendix E). The inner pipe-grout heat flux resistance relationships are shown in Fig. M.1.

M.1.2 Single U-Shape Pipe (1U)

The single U-shape pipe (1U) exchanger can be easily degenerated from a 2U configuration when dropping the second U-tube. A 1U configuration only consists of four components (Fig. M.2):

- One pipe-in (denoted as $i1$)
- One pipe-out (denoted as $o1$)
- Grout material, which is subdivided into two zones (denoted as $g1$, $g2$)

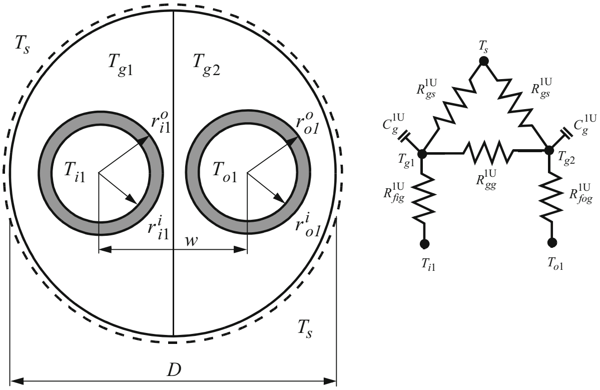


Fig. M.2 Inner pipe-grout heat flux resistance relationships of a 1U borehole consisting of two pipe components and two grout zones

Similar to the 2U exchanger the U-tube of the 1U configuration transfers heat in radial direction to the grout zones gi $(i = 1, \ldots, 2)$, while the grout material zones exchange heat directly to the surrounding soil s and to the adjacent grout zone. The corresponding inner pipe-grout heat flux resistance relationships are shown in Fig. M.2.

M.1.3 Coaxial Pipe with Annular (CXA) and Centered (CXC) Inlet

These types of BHE consist only of three components (Figs. M.3 and M.4):

- One pipe-in (denoted as $i1$)
- One pipe-out (denoted as $o1$)
- Grout material considered in one zone (denoted as $g1$)

Such coaxial BHE systems represent pipe-in-pipe installations, where two principal cases occur. In the case of the CXA exchanger the pipe-out is configured inside the pipe-in as shown in Fig. M.3 forming an annular inlet and a centered outlet. Accordingly, the heat exchange to the grout material $g1$, which is in contact to the surrounding soil s, is only performed via the pipe-in $i1$. On the other hand, the pipe-in $i1$ exchanges heat with the pipe-out $o1$ component. The coaxial pipes can also be installed with interchanged inlet and outlet. This represents the CXC exchanger, where the pipe-in is configured inside the pipe-out as shown in Fig. M.4 forming a centered inlet and an annular outlet. Here, the heat exchange to the grout material $g1$ is only performed via the pipe-out $o1$.

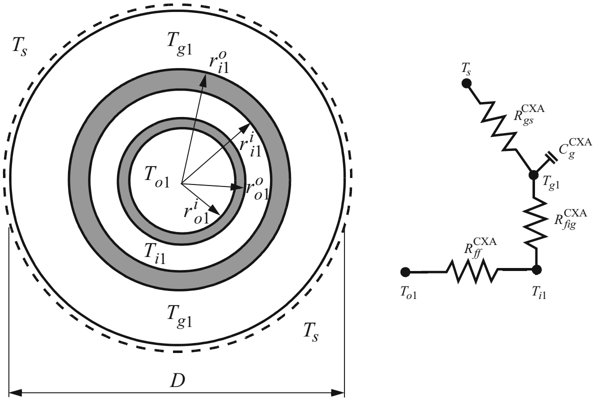


Fig. M.3 Inner pipe-grout heat flux resistance relationships of a CXA borehole with annular inlet

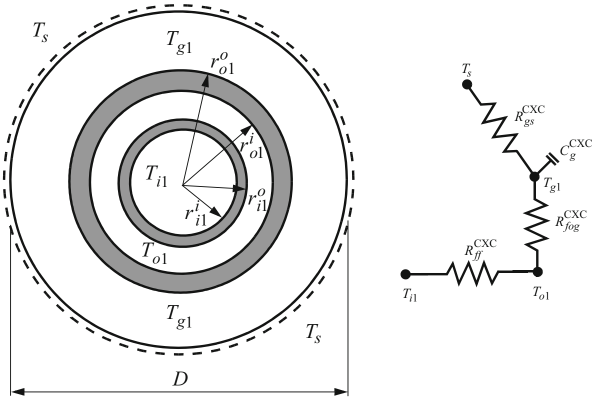


Fig. M.4 Inner pipe-grout heat flux resistance relationships of a CXC borehole with centered inlet

M.2 Thermal Resistances

Thermal resistance is a measure of material's ability to resist heat transfer through its surface and contact zone (cf. Appendix E). Thermal resistances are determined from the physical, material and geometric engineering parameters of the different BHE configurations as shown in Fig. M.1 for the 2U exchanger, in Fig. M.2 for the 1U exchanger, in Fig. M.3 for the CXA exchanger and in Fig. M.4 for the CXC exchanger. As indicated the interaction between the different components of the pipe exists between the pipe-in and grout zone(s), the pipe-out and grout zone(s) as well as the pipe-in and pipe-out. The thermal resistances due to heat conduction in the grout material are derived by adding correction terms gained from numerical simulations to well-known 2D heat conduction shape factors. These resistances then are divided in such a manner, that the grout points are suitably located to obtain accurate transient computation results, see [29]. The following *specific* thermal resistances can be derived.

M.2.1 2U Exchanger

The thermal resistance between the pipes and grout zones is caused by the advection of the pipe flow and thermal conductivity of the pipe wall material specified separately for pipe-in and pipe-out

$$R^{2U}_{fig} = R^{2U}_{\mathrm{adv}_k} + R^{2U}_{\mathrm{con}^a_k} + R^{2U}_{\mathrm{con}^b} \quad (k = i1 \cap i2) \tag{M.1}$$

$$R^{2U}_{fog} = R^{2U}_{\mathrm{adv}_k} + R^{2U}_{\mathrm{con}^a_k} + R^{2U}_{\mathrm{con}^b} \quad (k = o1 \cap o2) \tag{M.2}$$

M.2.1.1 Thermal Resistance Due to the Advective Flow of Refrigerant in the Pipes

$$R^{2U}_{\mathrm{adv}_k} = \frac{1}{\mathrm{Nu}_k \Lambda^r \pi} \quad (k = i1, o1, i2, o2) \tag{M.3}$$

where Λ^r is the thermal conductivity of the refrigerant. In (M.3) the Nusselt numbers, $\mathrm{Nu}_k (k = i1, o1, i2, o2)$, differ between laminar and turbulent flow [544], viz.,

$$\mathrm{Nu}_k = \begin{cases} 4.364 \\ \text{for laminar flow if } \mathrm{Re}_k < 2,300 \\ \frac{(\xi_k/8)\mathrm{Re}_k \mathrm{Pr}}{1+12.7\sqrt{\xi_k/8}(\mathrm{Pr}^{2/3}-1)}\left[1 + \left(\frac{d^i_k}{L}\right)^{2/3}\right] \\ \text{for turbulent flow if } \mathrm{Re}_k \geq 10^4 \\ (1-\gamma_k)4.364 + \gamma_k\left\{\frac{(0.0308/8)10^4\mathrm{Pr}}{1+12.7\sqrt{0.0308/8}(\mathrm{Pr}^{2/3}-1)}\left[1 + \left(\frac{d^i_k}{L}\right)^{2/3}\right]\right\} \\ \text{for flow in transition range if } 2{,}300 \leq \mathrm{Re}_k < 10^4 \end{cases} \tag{M.4}$$

in which Pr represents the Prandtl number and Re_k are the Reynolds numbers defined as

$$\mathrm{Pr} = \frac{\mu^r c^r}{\Lambda^r}, \quad \mathrm{Re}_k = \frac{|\boldsymbol{u}_k|^{2U} d^i_k}{(\mu^r/\rho^r)} \quad (k = i1, o1, i2, o2) \tag{M.5}$$

where d^i_k are the inner diameters of the pipes $d^i_k = 2r^i_k(k = i1, o1, i2, o2)$. Furthermore, L corresponds to the length of the pipe and

$$\left.\begin{aligned} \xi_k &= \left(1.8 \log_{10} \mathrm{Re}_k - 1.5\right)^{-2} \\ \gamma_k &= \tfrac{\mathrm{Re}_k - 2{,}300}{10^4 - 2{,}300} \qquad (0 \leq \gamma_k \leq 1) \end{aligned}\right\} \tag{M.6}$$

It is

$$|\boldsymbol{u}_k|^{2\mathrm{U}} = \begin{cases} \dfrac{Q_r}{2\pi (r_k^i)^2} & \text{for parallel discharge} \\ \dfrac{Q_r}{\pi (r_k^i)^2} & \text{for serial discharge} \end{cases} \qquad (k = i1, o1, i2, o2) \tag{M.7}$$

where Q_r is the total refrigerant flow discharge of the 2U exchanger.

M.2.1.2 Thermal Resistances Due to the Pipes Wall Material and Grout Transition

$$R^{2\mathrm{U}}_{\mathrm{con}^a_k} = \frac{\ln(r_k^o / r_k^i)}{2\pi \Lambda_k^\pi} \quad (k = i1, o1, i2, o2) \tag{M.8}$$

where Λ^π_{i1}, Λ^π_{o1}, Λ^π_{i2}, Λ^π_{o2} correspond to the thermal conductivities of the pipe wall material.

$$R^{2\mathrm{U}}_{\mathrm{con}^b} = x^{2\mathrm{U}} R_g^{2\mathrm{U}} \tag{M.9}$$

with

$$x^{2\mathrm{U}} = \frac{\ln\left(\frac{\sqrt{D^2+4d_o^2}}{2\sqrt{2}d_o}\right)}{\ln(\frac{D}{2d_o})} \tag{M.10}$$

and

$$R_g^{2\mathrm{U}} = \frac{\operatorname{arcosh}(\frac{D^2+d_o^2-s^2}{2Dd_o})}{2\pi \Lambda^g} \left(3.098 - 4.432\frac{s}{D} + 2.364\frac{s^2}{D^2}\right) \tag{M.11}$$

where D denotes the borehole diameter, $d_o = \frac{1}{4}\sum_k d_k^o$ is the average outer diameter of the pipes $d_k^o = 2r_k^o (k = i1, o1, i2, o2)$ and $s = w\sqrt{2}$ corresponds to diagonal distances of pipes (see Fig. M.1).

M.2.1.3 Thermal Resistance Due to Inter-grout Exchange

$$R^{2\mathrm{U}}_{gg1} = \frac{2R^{2\mathrm{U}}_{gs}(R^{2\mathrm{U}}_{ar1} - 2x^{2\mathrm{U}} R^{2\mathrm{U}}_g)}{2R^{2\mathrm{U}}_{gs} - R^{2\mathrm{U}}_{ar1} + 2x^{2\mathrm{U}} R^{2\mathrm{U}}_g} \tag{M.12}$$

$$R^{2\mathrm{U}}_{gg2} = \frac{2R^{2\mathrm{U}}_{gs}(R^{2\mathrm{U}}_{ar2} - 2x^{2\mathrm{U}} R^{2\mathrm{U}}_g)}{2R^{2\mathrm{U}}_{gs} - R^{2\mathrm{U}}_{ar2} + 2x^{2\mathrm{U}} R^{2\mathrm{U}}_g} \tag{M.13}$$

with

$$R_{ar1}^{2\mathrm{U}} = \frac{\operatorname{arcosh}(\frac{s^2-d_o^2}{d_o^2})}{2\pi \Lambda^g} \tag{M.14}$$

$$R_{ar2}^{2\mathrm{U}} = \frac{\operatorname{arcosh}(\frac{2s^2-d_o^2}{d_o^2})}{2\pi \Lambda^g} \tag{M.15}$$

M.2.1.4 Thermal Resistance Due to Grout-Soil Exchange

$$R_{gs}^{2\mathrm{U}} = (1 - x^{2\mathrm{U}}) R_g^{2\mathrm{U}} \tag{M.16}$$

M.2.2 1U Exchanger

It is

$$R_{\mathit{fig}}^{1\mathrm{U}} = R_{\mathrm{adv}_k}^{1\mathrm{U}} + R_{\mathrm{con}_k^a}^{1\mathrm{U}} + R_{\mathrm{con}^b}^{1\mathrm{U}} \quad (k = i1) \tag{M.17}$$

$$R_{\mathit{fog}}^{1\mathrm{U}} = R_{\mathrm{adv}_k}^{1\mathrm{U}} + R_{\mathrm{con}_k^a}^{1\mathrm{U}} + R_{\mathrm{con}^b}^{1\mathrm{U}} \quad (k = o1) \tag{M.18}$$

M.2.2.1 Thermal Resistance Due to the Advective Flow of Refrigerant in the Pipes

$$R_{\mathrm{adv}_k}^{1\mathrm{U}} = \frac{1}{\mathrm{Nu}_k \Lambda^r \pi} \quad (k = i1, o1) \tag{M.19}$$

where Nu_k is given by the expressions (M.4)–(M.6) in which the refrigerant fluid velocity for 1D pipe is

$$|\boldsymbol{u}_k|^{1\mathrm{U}} = \frac{Q_r}{2\pi (r_k^i)^2} \quad (k = i1, o1) \tag{M.20}$$

M.2.2.2 Thermal Resistance Due to the Pipes Wall Material and Grout Transition

$$R_{\mathrm{con}_k^a}^{1\mathrm{U}} = \frac{\ln(r_k^o / r_k^i)}{2\pi \Lambda_k^\pi} \quad (k = i1, o1) \tag{M.21}$$

$$R_{\mathrm{con}^b}^{1\mathrm{U}} = x^{1\mathrm{U}} R_g^{1\mathrm{U}} \tag{M.22}$$

with

$$x^{1\mathrm{U}} = \frac{\ln\left(\frac{\sqrt{D^2+2d_o^2}}{2d_o}\right)}{\ln(\frac{D}{\sqrt{2}d_o})} \tag{M.23}$$

and

$$R_g^{1\mathrm{U}} = \frac{\operatorname{arcosh}(\frac{D^2+d_o^2-w^2}{2Dd_o})}{2\pi \Lambda^g}\left(1.601 - 0.888\frac{w}{D}\right) \tag{M.24}$$

where w corresponds to distances between the pipes (see Fig. M.2).

M.2.2.3 Thermal Resistance Due to Inter-grout Exchange

$$R_{gg}^{1\mathrm{U}} = \frac{2R_{gs}^{1\mathrm{U}}(R_{ar}^{1\mathrm{U}} - 2x^{1\mathrm{U}}R_g^{1\mathrm{U}})}{2R_{gs}^{1\mathrm{U}} - R_{ar}^{1\mathrm{U}} + 2x^{1\mathrm{U}}R_g^{1\mathrm{U}}} \tag{M.25}$$

with

$$R_{ar}^{1\mathrm{U}} = \frac{\operatorname{arcosh}(\frac{2w^2-d_o^2}{d_o^2})}{2\pi \Lambda^g} \tag{M.26}$$

M.2.2.4 Thermal Resistance Due to Grout-Soil Exchange

$$R_{gs}^{1\mathrm{U}} = (1 - x^{1\mathrm{U}})R_g^{1\mathrm{U}} \tag{M.27}$$

M.2.3 CXA Exchanger

It is

$$R_{\mathit{ff}}^{\mathrm{CXA}} = R_{\mathrm{adv}_{o1}}^{\mathrm{CXA}} + R_{\mathrm{adv}_{i1}^{a}}^{\mathrm{CXA}} + R_{\mathrm{con}_{o1}}^{\mathrm{CXA}} \tag{M.28}$$

$$R_{\mathit{fig}}^{\mathrm{CXA}} = R_{\mathrm{adv}_{i1}^{b}}^{\mathrm{CXA}} + R_{\mathrm{con}_{i1}}^{\mathrm{CXA}} + R_{\mathrm{con}^{b}}^{\mathrm{CXA}} \tag{M.29}$$

M.2.3.1 Thermal Resistance Due to the Advective Flow of Refrigerant in the Pipes

$$R^{\text{CXA}}_{\text{adv}_{o1}} = \frac{1}{\text{Nu}_{o1}\Lambda^r\pi} \tag{M.30}$$

$$R^{\text{CXA}}_{\text{adv}^a_{i1}} = \frac{1}{\text{Nu}_{i1}\Lambda^r\pi}\frac{d_h}{d^o_{o1}} \tag{M.31}$$

$$R^{\text{CXA}}_{\text{adv}^b_{i1}} = \frac{1}{\text{Nu}_{i1}\Lambda^r\pi}\frac{d_h}{d^i_{i1}} \tag{M.32}$$

$$\text{Nu}_{o1} = \begin{cases} 4.364 \\ \text{for laminar flow if } \text{Re}_{o1} < 2,300 \\ \frac{(\xi_{o1}/8)\text{Re}_{o1}\text{Pr}}{1+12.7\sqrt{\xi_{o1}/8}(\text{Pr}^{2/3}-1)}\left[1+\left(\frac{d^i_{o1}}{L}\right)^{2/3}\right] \\ \text{for turbulent flow if } \text{Re}_{o1} \geq 10^4 \\ (1-\gamma_{o1})4.364+\gamma_{o1}\left\{\frac{(0.0308/8)10^4\text{Pr}}{1+12.7\sqrt{0.0308/8}(\text{Pr}^{2/3}-1)}\left[1+\left(\frac{d^i_{o1}}{L}\right)^{2/3}\right]\right\} \\ \text{for flow in transition range if } 2{,}300 \leq \text{Re}_{o1} < 10^4 \end{cases} \tag{M.33}$$

$$\text{Nu}_{i1} = \begin{cases} 3.66+\left[4-\frac{0.102}{\left(\frac{d^o_{o1}}{d^i_{i1}}\right)+0.02}\right]\left(\frac{d^o_{o1}}{d^i_{i1}}\right)^{0.04} \\ \text{for laminar flow if } \text{Re}_{i1} < 2,300 \\ \frac{(\xi_{i1}/8)\text{Re}_{i1}\text{Pr}}{1+12.7\sqrt{\xi_{i1}/8}(\text{Pr}^{2/3}-1)}\left[1+\left(\frac{d_h}{L}\right)^{2/3}\right]\left\{\frac{0.86\left(\frac{d^o_{o1}}{d^i_{i1}}\right)^{0.84}+\left[1-0.14\left(\frac{d^o_{o1}}{d^i_{i1}}\right)^{0.6}\right]}{1+\left(\frac{d^o_{o1}}{d^i_{i1}}\right)}\right\} \\ \text{for turbulent flow if } \text{Re}_{i1} \geq 10^4 \\ (1-\gamma_{i1})\left\{3.66+\left[4-\frac{0.102}{\left(\frac{d^o_{o1}}{d^i_{i1}}\right)+0.02}\right]\right\}\left(\frac{d^o_{o1}}{d^i_{i1}}\right)^{0.04}+ \\ \gamma_{i1}\left\{\frac{(0.0308/8)10^4\text{Pr}}{1+12.7\sqrt{0.0308/8}(\text{Pr}^{2/3}-1)}\times\right. \\ \quad\left.\left[1+\left(\frac{d_h}{L}\right)^{2/3}\right]\left\{\frac{0.86\left(\frac{d^o_{o1}}{d^i_{i1}}\right)^{0.84}+\left[1-0.14\left(\frac{d^o_{o1}}{d^i_{i1}}\right)^{0.6}\right]}{1+\left(\frac{d^o_{o1}}{d^i_{i1}}\right)}\right\}\right\} \\ \text{for flow in transition range if } 2{,}300 \leq \text{Re}_{i1} < 10^4 \end{cases} \tag{M.34}$$

where

$$\text{Pr} = \frac{\mu^r c^r}{\Lambda^r}, \quad \text{Re}_{o1} = \frac{|\boldsymbol{u}_{o1}|^{\text{CXA}}d^i_{o1}}{(\mu^r/\rho^r)}, \quad \text{Re}_{i1} = \frac{|\boldsymbol{u}_{11}|^{\text{CXA}}d_h}{(\mu^r/\rho^r)} \tag{M.35}$$

and

$$\left.\begin{aligned} d_h &= d_{i1}^i - d_{o1}^o \\ \xi_k &= \big(1.8\log_{10}\mathrm{Re}_k - 1.5\big)^{-2} \\ \gamma_k &= \frac{\mathrm{Re}_k - 2,300}{10^4 - 2,300} \qquad (0 \le \gamma_k \le 1) \\ &(k = i1, o1) \end{aligned}\right\} \tag{M.36}$$

$$\left.\begin{aligned} |\boldsymbol{u}_{o1}|^{\mathrm{CXA}} &= \frac{Q_r}{2\pi(r_{o1}^i)^2} \\ |\boldsymbol{u}_{i1}|^{\mathrm{CXA}} &= \frac{Q_r}{2\pi[(r_{i1}^i)^2 - (r_{o1}^o)^2]} \end{aligned}\right\} \tag{M.37}$$

M.2.3.2 Thermal Resistance Due to the Pipes Wall Material and Grout Transition

$$R_{\mathrm{con}_k}^{\mathrm{CXA}} = \frac{\ln(r_k^o/r_k^i)}{2\pi\Lambda_k^\pi} \quad (k = i1, o1) \tag{M.38}$$

$$R_{\mathrm{con}^b}^{\mathrm{CXA}} = x^{\mathrm{CXA}} R_g^{\mathrm{CXA}} \tag{M.39}$$

with

$$x^{\mathrm{CXA}} = \frac{\ln\left(\frac{\sqrt{D^2+(d_{i1}^o)^2}}{\sqrt{2}d_{i1}^o}\right)}{\ln(\frac{D}{d_{i1}^o})} \tag{M.40}$$

and

$$R_g^{\mathrm{CXA}} = \frac{\ln(D/d_{i1}^o)}{2\pi\Lambda^g} \tag{M.41}$$

M.2.3.3 Thermal Resistance Due to Grout-Soil Exchange

$$R_{gs}^{\mathrm{CXA}} = (1 - x^{\mathrm{CXA}})R_g^{\mathrm{CXA}} \tag{M.42}$$

M.2.4 CXC Exchanger

$$R_{ff}^{\mathrm{CXC}} = R_{\mathrm{adv}_{i1}}^{\mathrm{CXC}} + R_{\mathrm{adv}_{o1}^a}^{\mathrm{CXC}} + R_{\mathrm{con}_{i1}}^{\mathrm{CXC}} \tag{M.43}$$

$$R_{fog}^{\mathrm{CXC}} = R_{\mathrm{adv}_{o1}^b}^{\mathrm{CXC}} + R_{\mathrm{con}_{o1}}^{\mathrm{CXC}} + R_{\mathrm{con}^b}^{\mathrm{CXC}} \tag{M.44}$$

M.2.4.1 Thermal Resistance Due to the Advective Flow of Refrigerant in the Pipes

$$R^{\mathrm{CXC}}_{\mathrm{adv}_{i1}} = \frac{1}{\mathrm{Nu}_{i1} \Lambda^r \pi} \tag{M.45}$$

$$R^{\mathrm{CXC}}_{\mathrm{adv}^a_{o1}} = \frac{1}{\mathrm{Nu}_{o1} \Lambda^r \pi} \frac{d_h}{d^o_{i1}} \tag{M.46}$$

$$R^{\mathrm{CXC}}_{\mathrm{adv}^b_{o1}} = \frac{1}{\mathrm{Nu}_{o1} \Lambda^r \pi} \frac{d_h}{d^i_{o1}} \tag{M.47}$$

with

$$\mathrm{Nu}_{i1} = \begin{cases} 4.364 \\ \text{for laminar flow if } \mathrm{Re}_{i1} < 2,300 \\ \frac{(\xi_{i1}/8)\mathrm{Re}_{i1}\mathrm{Pr}}{1+12.7\sqrt{\xi_{i1}/8}(\mathrm{Pr}^{2/3}-1)}\left[1+\left(\frac{d^i_{i1}}{L}\right)^{2/3}\right] \\ \text{for turbulent flow if } \mathrm{Re}_{i1} \geq 10^4 \\ (1-\gamma_{i1})4.364 + \gamma_{i1}\left\{\frac{(0.0308/8)10^4\mathrm{Pr}}{1+12.7\sqrt{0.0308/8}(\mathrm{Pr}^{2/3}-1)}\left[1+\left(\frac{d^i_{i1}}{L}\right)^{2/3}\right]\right\} \\ \text{for flow in transition range if } 2{,}300 \leq \mathrm{Re}_{i1} < 10^4 \end{cases} \tag{M.48}$$

$$\mathrm{Nu}_{o1} = \begin{cases} 3.66 + \left[4 - \frac{0.102}{\left(\frac{d^o_{i1}}{d^i_{o1}}\right)+0.02}\right]\left(\frac{d^o_{i1}}{d^i_{o1}}\right)^{0.04} \\ \text{for laminar flow if } \mathrm{Re}_{o1} < 2,300 \\ \frac{(\xi_{o1}/8)\mathrm{Re}_{o1}\mathrm{Pr}}{1+12.7\sqrt{\xi_{o1}/8}(\mathrm{Pr}^{2/3}-1)}\left[1+\left(\frac{d_h}{L}\right)^{2/3}\right]\left\{\frac{0.86\left(\frac{d^o_{i1}}{d^i_{o1}}\right)^{0.84}+\left[1-0.14\left(\frac{d^o_{i1}}{d^i_{o1}}\right)^{0.6}\right]}{1+\left(\frac{d^o_{i1}}{d^i_{o1}}\right)}\right\} \\ \text{for turbulent flow if } \mathrm{Re}_{o1} \geq 10^4 \\ (1-\gamma_{o1})\left\{3.66 + \left[4 - \frac{0.102}{\left(\frac{d^o_{i1}}{d^i_{o1}}\right)+0.02}\right]\right\}\left(\frac{d^o_{i1}}{d^i_{o1}}\right)^{0.04} + \\ \gamma_{o1}\left\{\frac{(0.0308/8)10^4\mathrm{Pr}}{1+12.7\sqrt{0.0308/8}(\mathrm{Pr}^{2/3}-1)} \times \right. \\ \left.\left[1+\left(\frac{d_h}{L}\right)^{2/3}\right]\left\{\frac{0.86\left(\frac{d^o_{i1}}{d^i_{o1}}\right)^{0.84}+\left[1-0.14\left(\frac{d^o_{i1}}{d^i_{o1}}\right)^{0.6}\right]}{1+\left(\frac{d^o_{i1}}{d^i_{o1}}\right)}\right\}\right\} \\ \text{for flow in transition range if } 2{,}300 \leq \mathrm{Re}_{o1} < 10^4 \end{cases} \tag{M.49}$$

where

$$\mathrm{Pr} = \frac{\mu^r c^r}{\Lambda^r} \quad \mathrm{Re}_{i1} = \frac{|\boldsymbol{u}_{i1}|^{\mathrm{CXC}} d_{i1}^i}{(\mu^r/\rho^r)} \quad \mathrm{Re}_{o1} = \frac{|\boldsymbol{u}_{o1}|^{\mathrm{CXC}} d_h}{(\mu^r/\rho^r)} \tag{M.50}$$

and

$$\left.\begin{array}{l} d_h = d_{o1}^i - d_{i1}^o \\ \xi_k = \left(1.8\log_{10}\mathrm{Re}_k - 1.5\right)^{-2} \\ \gamma_k = \dfrac{\mathrm{Re}_k - 2,300}{10^4 - 2,300} \qquad (0 \le \gamma_k \le 1) \\ \quad (k = i1, o1) \end{array}\right\} \tag{M.51}$$

$$\left.\begin{array}{l} |\boldsymbol{u}_{i1}|^{\mathrm{CXC}} = \dfrac{Q_r}{2\pi(r_{i1}^i)^2} \\ |\boldsymbol{u}_{o1}|^{\mathrm{CXC}} = \dfrac{Q_r}{2\pi[(r_{o1}^i)^2 - (r_{i1}^o)^2]} \end{array}\right\} \tag{M.52}$$

M.2.4.2 Thermal Resistance Due to the Pipes Wall Material and Grout Transition

$$R_{\mathrm{con}_k}^{\mathrm{CXC}} = \frac{\ln(r_k^o/r_k^i)}{2\pi\Lambda_k^\pi} \quad (k = i1, o1) \tag{M.53}$$

$$R_{\mathrm{con}^b}^{\mathrm{CXC}} = x^{\mathrm{CXC}} R_g^{\mathrm{CXC}} \tag{M.54}$$

with

$$x^{\mathrm{CXC}} = \frac{\ln\left(\frac{\sqrt{D^2+(d_{o1}^o)^2}}{\sqrt{2}d_{o1}^o}\right)}{\ln(\frac{D}{d_{o1}^o})} \tag{M.55}$$

and

$$R_g^{\mathrm{CXC}} = \frac{\ln(D/d_{o1}^o)}{2\pi\Lambda^g} \tag{M.56}$$

M.2.4.3 Thermal Resistance Due to Grout-Soil Exchange

$$R_{gs}^{\mathrm{CXC}} = (1 - x^{\mathrm{CXC}}) R_g^{\mathrm{CXC}} \tag{M.57}$$

M.2.5 *Notes to Negative Thermal Resistances of Grout for 2U and 1U Exchangers*

In dependence on geometric measures for 2U and 1U exchangers negative thermal resistances for grout R_{gg1}^{2U}, R_{gg2}^{2U}, R_{gg}^{1U} may occur. This is caused by the applied TRCM conception of grout zones and can be accepted in both numerical and analytical BHE models. However, the following constraints have to be satisfied:

$$\begin{aligned} \left(\frac{1}{R_{gg1}^{2U}} + \frac{1}{2R_{gs}^{2U}}\right)^{-1} &> 0 \\ \left(\frac{1}{R_{gg2}^{2U}} + \frac{1}{2R_{gs}^{2U}}\right)^{-1} &> 0 \end{aligned} \tag{M.58}$$

for 2U exchangers and

$$\left(\frac{1}{R_{gg}^{1U}} + \frac{1}{2R_{gs}^{1U}}\right)^{-1} > 0 \tag{M.59}$$

for 1U exchangers. In cases where (M.58) or (M.59) are violated the values of x^{2U} and x^{1U}, respectively, have to be reduced until the constraints (M.58) and (M.59) are met. The following correction procedure is applied:

1. If (M.58) or (M.59) are violated reduce $x_{\text{new}}^{2U,1U} = \frac{2}{3}x_{\text{old}}^{2U,1U}$ and check (M.58) or (M.59).
2. If (M.58) or (M.59) are still violated reduce $x_{\text{new}}^{2U,1U} = \frac{1}{3}x_{\text{old}}^{2U,1U}$ and check (M.58) or (M.59).
3. If (M.58) or (M.59) are again violated set $x_{\text{new}}^{2U,1U} = 0$.

M.2.6 *Direct Use of Borehole Thermal Resistances R_a and R_b Obtained from Thermal Response Test (TRT)*

From practical point of view it could be useful to specify directly thermal resistances which have been measured in the field. Such field-related thermal resistances result for instance from Thermal Response Tests (TRT's), e.g., [25]. In such cases the *borehole thermal resistance* R_b and the *internal borehole thermal resistance* R_a are determined according to the definition introduced by Hellström [237] given for a Delta configuration of thermal circuit such as described in Sect. E.3.1 of Appendix E. With known R_b and R_a the complete set of thermal BHE resistances can be determined in dependence on the chosen analytical or numerical solution strategy as follows.

M.2.6.1 Analytical BHE Solution

2U Exchanger:

Replace (M.1), (M.2) and (M.16) by

$$\begin{aligned} R_{fig}^{2\mathrm{U}} &= 2R_b \\ R_{fog}^{2\mathrm{U}} &= 2R_b \\ R_{gs}^{2\mathrm{U}} &= 2R_b \end{aligned} \tag{M.60}$$

respectively, as well as (M.12) and (M.13) by

$$\begin{aligned} R_{gg1}^{2\mathrm{U}} &= \frac{8R_b(R_a - 2R_b)}{4R_b - R_a} \\ R_{gg2}^{2\mathrm{U}} &= R_{gg1}^{2\mathrm{U}} \end{aligned} \tag{M.61}$$

respectively.

1U Exchanger:

Replace (M.17), (M.18), and (M.27) by

$$\begin{aligned} R_{fig}^{1\mathrm{U}} &= R_b \\ R_{fog}^{1\mathrm{U}} &= R_b \\ R_{gs}^{1\mathrm{U}} &= R_b \end{aligned} \tag{M.62}$$

respectively, as well as (M.25) by

$$R_{gg}^{1\mathrm{U}} = \frac{2R_b(R_a - 2R_b)}{4R_b - R_a} \tag{M.63}$$

CXA Exchanger:

Replace (M.28), (M.29), and (M.42) by

$$\begin{aligned} R_{ff}^{\mathrm{CXA}} &= R_a \\ R_{fig}^{\mathrm{CXA}} &= \tfrac{R_b}{2} \\ R_{gs}^{\mathrm{CXA}} &= \tfrac{R_b}{2} \end{aligned} \tag{M.64}$$

respectively.

CXC Exchanger:

Replace (M.43), (M.44), and (M.57) by

$$\begin{aligned} R_{ff}^{\mathrm{CXC}} &= R_a \\ R_{fog}^{\mathrm{CXC}} &= \tfrac{R_b}{2} \\ R_{gs}^{\mathrm{CXC}} &= \tfrac{R_b}{2} \end{aligned} \tag{M.65}$$

respectively.

M.2.6.2 Numerical BHE Solution

2U Exchanger:

Defining

$$\begin{aligned} R_{\mathrm{adv}}^{2\mathrm{U}} &= \tfrac{1}{4}(R_{\mathrm{adv}_{i1}}^{2\mathrm{U}} + R_{\mathrm{adv}_{i2}}^{2\mathrm{U}} + R_{\mathrm{adv}_{o1}}^{2\mathrm{U}} + R_{\mathrm{adv}_{o2}}^{2\mathrm{U}}) \\ R_{\mathrm{con}}^{2\mathrm{U}} &= \tfrac{1}{4}(R_{\mathrm{con}_{i1}^a}^{2\mathrm{U}} + R_{\mathrm{con}_{i2}^a}^{2\mathrm{U}} + R_{\mathrm{con}_{o1}^a}^{2\mathrm{U}} + R_{\mathrm{con}_{o2}^a}^{2\mathrm{U}}) \end{aligned} \tag{M.66}$$

we replace (M.11), (M.14), and (M.15) by

$$\begin{aligned} R_g^{2\mathrm{U}} &= 4R_b - R_{\mathrm{adv}}^{2\mathrm{U}} - R_{\mathrm{con}}^{2\mathrm{U}} \\ R_{ar1}^{2\mathrm{U}} &= \frac{(2+\sqrt{2})R_g^{2\mathrm{U}}(R_a - R_{\mathrm{adv}}^{2\mathrm{U}} - R_{\mathrm{con}}^{2\mathrm{U}})}{R_g^{2\mathrm{U}} + R_a - R_{\mathrm{adv}}^{2\mathrm{U}} - R_{\mathrm{con}}^{2\mathrm{U}}} \\ R_{ar2}^{2\mathrm{U}} &= \sqrt{2}R_{ar1}^{2\mathrm{U}} \end{aligned} \tag{M.67}$$

respectively.

1U Exchanger:

Defining

$$\begin{aligned} R_{\mathrm{adv}}^{1\mathrm{U}} &= \tfrac{1}{2}(R_{\mathrm{adv}_{i1}}^{1\mathrm{U}} + R_{\mathrm{adv}_{o1}}^{1\mathrm{U}}) \\ R_{\mathrm{con}}^{1\mathrm{U}} &= \tfrac{1}{2}(R_{\mathrm{con}_{i1}^a}^{1\mathrm{U}} + R_{\mathrm{con}_{o1}^a}^{1\mathrm{U}}) \end{aligned} \tag{M.68}$$

we replace (M.24) and (M.26) by

$$\begin{aligned} R_g^{1\mathrm{U}} &= 2R_b - R_{\mathrm{adv}}^{1\mathrm{U}} - R_{\mathrm{con}}^{1\mathrm{U}} \\ R_{ar}^{1\mathrm{U}} &= R_a - 2(R_{\mathrm{adv}}^{1\mathrm{U}} + R_{\mathrm{con}}^{1\mathrm{U}}) \end{aligned} \tag{M.69}$$

respectively.

CXA Exchanger:

Replace (M.28) and (M.41) by

$$\begin{aligned} R_{ff}^{\mathrm{CXA}} &= R_a \\ R_g^{\mathrm{CXA}} &= R_b - R_{\mathrm{adv}_{i1}^b}^{\mathrm{CXA}} - R_{\mathrm{con}_{i1}}^{\mathrm{CXA}} \end{aligned} \tag{M.70}$$

respectively.

CXC Exchanger:

Replace (M.43) and (M.56) by

$$\begin{aligned} R_{ff}^{\mathrm{CXC}} &= R_a \\ R_g^{\mathrm{CXC}} &= R_b - R_{\mathrm{adv}_{o1}^b}^{\mathrm{CXC}} - R_{\mathrm{con}_{o1}}^{\mathrm{CXC}} \end{aligned} \tag{M.71}$$

respectively.

M.3 Heat Transfer Coefficients

A heat transfer coefficient Φ_T represents the reciprocal of the effective specific thermal resistance R with its specific exchange area S (cf. Sect. E.2 of Appendix E), viz.,

$$\Phi_T = \frac{1}{R\,S} \tag{M.72}$$

Accordingly, the specific thermal resistances R_{fig}, R_{fog}, R_{gg1}, R_{gg2}, R_{ff}, R_{gg} and R_{gs} given by the relationships in the preceding Sect. M.2 for the 2U, 1U, CXA and CXC configurations can be used to express the corresponding heat transfer coefficients Φ_{fig}, Φ_{fog}, Φ_{gg1}, Φ_{gg2}, Φ_{ff}, Φ_{gg} and Φ_{gs} for the BHE configurations as follows:

2U Exchanger:

$$\left.\begin{aligned} \Phi_{fig}^{2\mathrm{U}} &= \frac{1}{R_{fig}^{2\mathrm{U}}}\frac{1}{S_i} \\ \Phi_{fog}^{2\mathrm{U}} &= \frac{1}{R_{fog}^{2\mathrm{U}}}\frac{1}{S_o} \\ \Phi_{gg1}^{2\mathrm{U}} &= \frac{1}{R_{gg1}^{2\mathrm{U}}}\frac{1}{S_{g1}} \\ \Phi_{gg2}^{2\mathrm{U}} &= \frac{1}{R_{gg2}^{2\mathrm{U}}}\frac{1}{S_{g2}} \\ \Phi_{gs}^{2\mathrm{U}} &= \frac{1}{R_{gs}^{2\mathrm{U}}}\frac{1}{S_{gs}} \end{aligned}\right\} \tag{M.73}$$

Table M.1 Specific exchange surfaces S for the BHE configurations

S	2U	1U	CXA	CXC
S_i	$\pi d^i_{i1,i2}$	πd^i_{i1}	πd^i_{i1}	–
S_o	$\pi d^i_{o1,o2}$	πd^i_{o1}	–	πd^i_{o1}
S_{io}	–	–	πd^i_{o1}	πd^i_{i1}
S_{g1}	$\frac{1}{2}D$	D	–	–
S_{g2}	D	–	–	–
S_{gs}	$\frac{1}{4}\pi D$	$\frac{1}{4}\pi D$	πD	πD

1U Exchanger:

$$\left.\begin{aligned} \Phi^{1U}_{fig} &= \frac{1}{R^{1U}_{fig}}\frac{1}{S_i} \\ \Phi^{1U}_{fog} &= \frac{1}{R^{1U}_{fog}}\frac{1}{S_o} \\ \Phi^{1U}_{gg} &= \frac{1}{R^{1U}_{gg}}\frac{1}{S_{g1}} \\ \Phi^{1U}_{gs} &= \frac{1}{R^{1U}_{gs}}\frac{1}{S_{gs}} \end{aligned}\right\} \tag{M.74}$$

CXA Exchanger:

$$\left.\begin{aligned} \Phi^{CXA}_{fig} &= \frac{1}{R^{CXA}_{fig}}\frac{1}{S_i} \\ \Phi^{CXA}_{ff} &= \frac{1}{R^{CXA}_{ff}}\frac{1}{S_{io}} \\ \Phi^{CXA}_{gs} &= \frac{1}{R^{CXA}_{gs}}\frac{1}{S_{gs}} \end{aligned}\right\} \tag{M.75}$$

CXC Exchanger:

$$\left.\begin{aligned} \Phi^{CXC}_{fog} &= \frac{1}{R^{CXC}_{fog}}\frac{1}{S_o} \\ \Phi^{CXC}_{ff} &= \frac{1}{R^{CXC}_{ff}}\frac{1}{S_{io}} \\ \Phi^{CXC}_{gs} &= \frac{1}{R^{CXC}_{gs}}\frac{1}{S_{gs}} \end{aligned}\right\} \tag{M.76}$$

where the related specific exchange areas of the BHE configurations are listed in Table M.1.

M.4 Analytical Solution of the Local Problem

In this section, the BHE equations are solved by Eskilson and Claesson's analytical method [159]. This method is applied to the different BHE configurations by using the extended TRCM [29]. Explicit relations result for temperatures of the pipes and the grout zones at vertical position and time. They are coupled to the soil temperature due to the incorporated heat transfer between BHE and the soil system. The analytical method has shown highly efficient, precise and robust, however, not applicable to short-term processes (a temporal scale in order of seconds, minutes or few hours).

M.4.1 Local Steady-State Condition with Given Temperature at Borehole Wall

The present analytical solution is only valid for local steady-state heat transport and given temperature $T_s = T_s(z,t)$ at borehole wall. It was firstly derived by Eskilson and Claesson [159] for heat transfer between two pipes and the borehole wall assuming a Delta configuration of the corresponding three thermal resistors with their specific thermal resistances R_1^{Δ}, R_2^{Δ} and R_{12}^{Δ} as illustrated in Fig. M.5, see also Sect. E.3.1 of Appendix E. We extend this analytical method to 2U, 1U, CXA and CXC configurations of BHE. In accordance with Fig. M.5, the local steady-state heat balance equations for fluid in pipe-in and pipe-out read

$$\begin{aligned} A^i \rho^r c^r |u| \frac{\partial T_{i1}}{\partial z} &= q_1 - q_{12} = \frac{T_s - T_{i1}}{R_1^{\Delta}} - \frac{T_{i1} - T_{o1}}{R_{12}^{\Delta}} \\ -A^i \rho^r c^r |u| \frac{\partial T_{o1}}{\partial z} &= q_2 + q_{12} = \frac{T_s - T_{o1}}{R_2^{\Delta}} + \frac{T_{i1} - T_{o1}}{R_{12}^{\Delta}} \end{aligned} \tag{M.77}$$

which have to be solved for the pipe(s)-in temperature $T_{i1}(z)$ and pipe(s)-out temperature $T_{o1}(z)$, where the $z-$coordinate is directed downward, superscript r refers to the refrigerant fluid, u is the refrigerant fluid velocity directed positive downward in $z-$direction (i.e., positive in pipe-in $i1$ and negative in pipe-out $o1$). They represent 1D heat transport equations for each pipe in the vertical $z-$direction, where the vertical heat conduction within the pipes is neglected. It is further assumed that the inner cross-sectional area of pipe-in and pipe-out is equal, i.e., $A^i = A_i^i = A_o^i$. The local steady-state condition limits the application of (M.77) to a time scale larger than [159]

$$t > t_{\text{limit}}^{\text{steady}} = \frac{5}{4} D^2 \Big(\frac{\varepsilon \rho c + (1-\varepsilon)\rho^s c^s}{\varepsilon \Lambda + (1-\varepsilon)\Lambda^s} \Big) \tag{M.78}$$

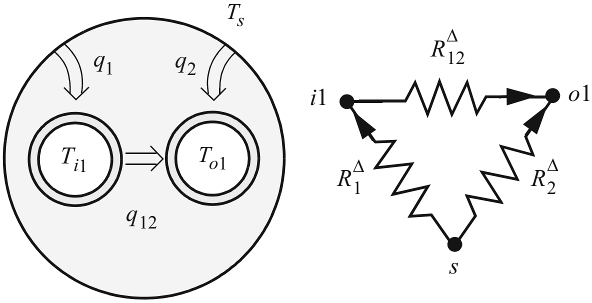


Fig. M.5 Borehole cross section, heat flow components and thermal circuit of Delta configuration (Modified from [159])

The time for the refrigerant to circulate through the borehole is $2A^i L/Q_r$. Accordingly, Eq. (M.77) can only describe transient input variations of inlet temperature and pumping rate on a time scale larger than [159]

$$t > t^{\text{steady}}_{\text{limit}} + A^i \frac{2L}{Q_r} \tag{M.79}$$

The specific thermal flux $\phi(z,t)$ exchanging heat of the borehole with the adjacent soil s is given from (M.77) according to

$$\phi(z,t) = \frac{T_s - T_{i1}}{R^{\vartriangle}_1} + \frac{T_s - T_{o1}}{R^{\vartriangle}_2} \tag{M.80}$$

M.4.2 Eskilson and Claesson's Analytical BHE Solution

The coupled equations (M.77) can be solved by using Laplace transforms [159]. It yields

$$\begin{aligned} T_{i1}(z,t) &= T_{i1}(0,t) f_1(z) + T_{o1}(0,t) f_2(z) + \int_0^z T_s(\xi,t) f_4(z-\xi) d\xi \\ & \qquad\qquad (0 \le z \le L) \\ T_{o1}(z,t) &= -T_{i1}(0,t) f_2(z) + T_{o1}(0,t) f_3(z) - \int_0^z T_s(\xi,t) f_5(z-\xi) d\xi \end{aligned} \tag{M.81}$$

The functions $f_1, f_2, \ldots, f_5$ are given by the expressions

$$\begin{aligned} f_1(z) &= e^{\beta z}[\cosh(\gamma z) - \delta \sinh(\gamma z)] \\ f_2(z) &= e^{\beta z} \frac{\beta_{12}}{\gamma} \sinh(\gamma z) \\ f_3(z) &= e^{\beta z}[\cosh(\gamma z) + \delta \sinh(\gamma z)] \end{aligned} \tag{M.82}$$

$$f_4(z) = e^{\beta z}\big[\beta_1 \cosh(\gamma z) - (\delta\beta_1 + \frac{\beta_2\beta_{12}}{\gamma}) \sinh(\gamma z)\big]$$

$$f_5(z) = e^{\beta z}\big[\beta_2 \cosh(\gamma z) + (\delta\beta_2 + \frac{\beta_1\beta_{12}}{\gamma}) \sinh(\gamma z)\big]$$

where

$$\beta_1 = \frac{1}{R_1^{\vartriangle} A^i \rho^r c^r u} \quad \beta_2 = \frac{1}{R_2^{\vartriangle} A^i \rho^r c^r u} \quad \beta_{12} = \frac{1}{R_{12}^{\vartriangle} A^i \rho^r c^r u} \quad \beta = \frac{\beta_2 - \beta_1}{2}$$
$$\gamma = \sqrt{\frac{(\beta_1 + \beta_2)^2}{4} + \beta_{12}(\beta_1 + \beta_2)} \quad \delta = \frac{1}{\gamma}\Big(\beta_{12} + \frac{\beta_1 + \beta_2}{2}\Big) \tag{M.83}$$

The following BC's are applied

$$\begin{aligned} T_{i1}(0, t) &= T_i(t) \\ T_{i1}(L, t) &= T_{o1}(L, t) \end{aligned} \tag{M.84}$$

where $T_i(t)$ represents the inlet temperature. Using (M.84) in (M.82) and (M.83) the outlet temperature $T_o(t)$ is given as

$$T_o(t) = T_{o1}(0, t) \tag{M.85}$$

M.4.3 Solutions for 1U and 2U Configurations

It is assumed that the pipes are arranged symmetrically within the borehole, i.e.,

$$R_2^{\vartriangle} = R_1^{\vartriangle} \tag{M.86}$$

so that

$$\left.\begin{aligned} \beta_2 = \beta_1 &= \frac{1}{R_1^{\vartriangle} A^i \rho^r c^r u} \\ \beta_{12} &= \frac{1}{R_{12}^{\vartriangle} A^i \rho^r c^r u} \\ \beta &= 0 \\ \gamma &= \sqrt{\beta_1^2 + 2\beta_{12}\beta_1} \\ \delta &= \tfrac{1}{\gamma}(\beta_{12} + \beta_1) \end{aligned}\right\} \tag{M.87}$$

Hence, (M.82) simplifies

$$
\begin{aligned}
f_1(z) &= \cosh(\gamma z) - \delta \sinh(\gamma z) \\
f_2(z) &= \frac{\beta_{12}}{\gamma} \sinh(\gamma z) \\
f_3(z) &= \cosh(\gamma z) + \delta \sinh(\gamma z) \\
f_4(z) &= \beta_1 \cosh(\gamma z) - (\delta\beta_1 + \frac{\beta_2\beta_{12}}{\gamma}) \sinh(\gamma z) \\
f_5(z) &= \beta_2 \cosh(\gamma z) + (\delta\beta_2 + \frac{\beta_1\beta_{12}}{\gamma}) \sinh(\gamma z)
\end{aligned}
\tag{M.88}
$$

In using (M.84) the Eq. (M.81) can be equalized at $z = L$ and solved for the outlet temperature $T_o(t)$, viz.,

$$
T_o(t) = T_i(t)\frac{f_1(L) + f_2(L)}{f_3(L) - f_2(L)} + \int_0^L \frac{T_s(\xi, t)[f_4(L-\xi) + f_5(L-\xi)]}{f_3(L) - f_2(L)} d\xi
\tag{M.89}
$$

With known inlet temperature $T_i(t)$ from the BC (M.84) and outlet temperature $T_o(t)$ from (M.89) the temperature distributions T_{i1} and T_{o1} as a function of z and t are obtained after evaluating the integrals in (M.81). It yields[1]

$$
\begin{aligned}
T_{i1}(z,t) &= T_i(t) f_1(z) + T_o(t) f_2(z) + \int_0^z T_s(\xi, t) f_4(z-\xi) d\xi \\
T_{o1}(z,t) &= -T_i(t) f_2(z) + T_o(t) f_3(z) - \int_0^z T_s(\xi, t) f_5(z-\xi) d\xi
\end{aligned}
\tag{M.90}
$$

Note that for the 2U configuration we assume $T_{i2} = T_{i1}$ and $T_{o2} = T_{o1}$. The integrals in (M.90) are performed elementwise, where the solid temperature T_s at the borehole wall is numerically approximated as a linear function from the nodal finite element solution at time t. For example

$$
\int_0^z T_s(\xi, t) f_4(z-\xi) d\xi \approx \sum_{e \in (z_1^e, z_2^e)} \frac{T_s^e(z_1^e, t) + T_s^e(z_2^e, t)}{2} F_4(z, z_2^e, z_1^e)
\tag{M.91}
$$

[1] The integrals $f_4(z-\xi)$ and $f_5(z-\xi)$ result for 1U and 2U configurations, respectively,

$$
\begin{aligned}
F_4(z,a,b) &= \int_a^b f_4(z-\xi) d\xi = -\tfrac{\beta_1}{\gamma} \sinh(\gamma(z-\xi))|_a^b + \left(\tfrac{\delta\beta_1}{\gamma} + \tfrac{\beta_2\beta_{12}}{\gamma^2}\right) \cosh(\gamma(z-\xi))|_a^b \\
F_5(z,a,b) &= \int_a^b f_5(z-\xi) d\xi = -\tfrac{\beta_2}{\gamma} \sinh(\gamma(z-\xi))|_a^b - \left(\tfrac{\delta\beta_2}{\gamma} + \tfrac{\beta_1\beta_{12}}{\gamma^2}\right) \cosh(\gamma(z-\xi))|_a^b
\end{aligned}
$$

where z_1^e, z_2^e represent the vertical coordinates of the lower and upper nodes, respectively, of element e.

The temperature distributions for the grout zones are derived from horizontal steady-state heat flow balances at the grout points, where the total heat exchange between the grout zones, pipe(s) and soil is assumed in equilibrium, such that

$$\begin{aligned} \frac{T_{g1}-T_s}{R_{gs}^{1\text{U}}} + \frac{T_{g1}-T_{i1}}{R_{fig}^{1\text{U}}} + \frac{T_{g1}-T_{g2}}{R_{gg}^{1\text{U}}} &= 0 \\ \frac{T_{g2}-T_s}{R_{gs}^{1\text{U}}} + \frac{T_{g2}-T_{o1}}{R_{fog}^{1\text{U}}} + \frac{T_{g2}-T_{g1}}{R_{gg}^{1\text{U}}} &= 0 \end{aligned} \tag{M.92}$$

given for the 1U exchanger (Fig. M.2) and

$$\begin{aligned} \frac{T_{g1}-T_s}{R_{gs}^{2\text{U}}} + \frac{T_{g1}-T_{i1}}{R_{fig}^{2\text{U}}} + \frac{T_{g1}-T_{g2}}{R_{gg2}^{2\text{U}}} + \frac{T_{g1}-T_{g3}}{R_{gg1}^{2\text{U}}} + \frac{T_{g1}-T_{g4}}{R_{gg1}^{2\text{U}}} &= 0 \\ \frac{T_{g2}-T_s}{R_{gs}^{2\text{U}}} + \frac{T_{g2}-T_{i2}}{R_{fig}^{2\text{U}}} + \frac{T_{g2}-T_{g1}}{R_{gg2}^{2\text{U}}} + \frac{T_{g2}-T_{g3}}{R_{gg1}^{2\text{U}}} + \frac{T_{g2}-T_{g4}}{R_{gg1}^{2\text{U}}} &= 0 \\ \frac{T_{g3}-T_s}{R_{gs}^{2\text{U}}} + \frac{T_{g3}-T_{o1}}{R_{fog}^{2\text{U}}} + \frac{T_{g3}-T_{g4}}{R_{gg2}^{2\text{U}}} + \frac{T_{g3}-T_{g1}}{R_{gg1}^{2\text{U}}} + \frac{T_{g3}-T_{g2}}{R_{gg1}^{2\text{U}}} &= 0 \\ \frac{T_{g4}-T_s}{R_{gs}^{2\text{U}}} + \frac{T_{g4}-T_{o2}}{R_{fog}^{2\text{U}}} + \frac{T_{g4}-T_{g3}}{R_{gg2}^{2\text{U}}} + \frac{T_{g4}-T_{g1}}{R_{gg1}^{2\text{U}}} + \frac{T_{g4}-T_{g2}}{R_{gg1}^{2\text{U}}} &= 0 \end{aligned} \tag{M.93}$$

given for the 2U exchanger (Fig. M.1). Accordingly, the temperature distribution for the two grout zones $T_{g1}(z,t)$ and $T_{g2}(z,t)$ of the 1U configuration can be derived as

$$\begin{aligned} T_{g1}(z,t) &= \frac{\left[\dfrac{T_s(z,t)}{R_{gs}^{1\text{U}}} + \dfrac{T_{o1}(z,t)}{R_{fog}^{1\text{U}}} + \left(\dfrac{T_s(z,t)}{R_{gs}^{1\text{U}}} + \dfrac{T_{i1}(z,t)}{R_{fig}^{1\text{U}}}\right) u_1 R_{gg}^{1\text{U}}\right] R_{gg}^{1\text{U}}}{(R_{gg}^{1\text{U}})^2 u_1^2 - 1} \\ T_{g2}(z,t) &= \left(\frac{T_{g1}(z,t)}{R_{gg}^{1\text{U}}} + \frac{T_{o1}(z,t)}{R_{fog}^{1\text{U}}} + \frac{T_s(z,t)}{R_{gs}^{1\text{U}}}\right)\frac{1}{u_1} \end{aligned} \tag{M.94}$$

with

$$u_1 = \frac{1}{R_{fig}^{1\text{U}}} + \frac{1}{R_{gs}^{1\text{U}}} + \frac{1}{R_{gg}^{1\text{U}}} \tag{M.95}$$

and the temperature distribution for the four grout zones $T_{g1}(z,t)$, $T_{g2}(z,t)$, $T_{g3}(z,t)$ and $T_{g4}(z,t)$ of the 2U configuration results in

$$T_{g1}(z,t) = T_{g2}(z,t) = \Big[\frac{2T_s(z,t)}{R_{gs}^{2\text{U}}} + \frac{2T_{o1}(z,t)}{R_{fog}^{2\text{U}}} + \Big(\frac{2T_s(z,t)}{R_{gs}^{2\text{U}}} + \frac{2T_{i1}(z,t)}{R_{fig}^{2\text{U}}}\Big)u_2 v\Big]\Big(\frac{v}{v^2u_2^2 - 1}\Big)$$

$$T_{g3}(z,t) = T_{g4}(z,t) = \Big(\frac{T_{g1}(z,t)}{v} + \frac{2T_{o1}(z,t)}{R_{fog}^{2\text{U}}} + \frac{2T_s(z,t)}{R_{gs}^{2\text{U}}}\Big)\frac{1}{u_2} \tag{M.96}$$

with

$$u_2 = \frac{2}{R_{fig}^{2\text{U}}} + \frac{2}{R_{gs}^{2\text{U}}} + \frac{1}{v}$$

$$v = \frac{R_{gg1}^{2\text{U}} R_{gg2}^{2\text{U}}}{2(R_{gg1}^{2\text{U}} + R_{gg2}^{2\text{U}})} \tag{M.97}$$

assuming $R_{fig}^{1\text{U}} = R_{fog}^{1\text{U}}$ and $R_{fig}^{2\text{U}} = R_{fog}^{2\text{U}}$. The thermal resistances $R_1^{\vartriangle}$ and $R_{12}^{\vartriangle}$ are given by

$$\left.\begin{aligned} R_1^{\vartriangle} &= R_{fig}^{1\text{U}} + R_{gs}^{1\text{U}} \\ R_{12}^{\vartriangle} &= \frac{(u_1 R_{fig}^{1\text{U}} R_{gg}^{1\text{U}})^2 - (R_{fig}^{1\text{U}})^2}{R_{gg}^{1\text{U}}} \end{aligned}\right\} \text{for 1U configuration} \tag{M.98}$$

and

$$\left.\begin{aligned} R_1^{\vartriangle} &= \frac{R_{fig}^{2\text{U}} + R_{gs}^{2\text{U}}}{2} \\ R_{12}^{\vartriangle} &= \frac{(R_{fig}^{2\text{U}})^2}{4}(u_2^2 v - \frac{1}{v}) \end{aligned}\right\} \text{for 2U configuration} \tag{M.99}$$

M.4.4 Solution for CXA Configuration

For coaxial BHE pipes with annular inlet there is

$$R_2^{\vartriangle} = \infty \tag{M.100}$$

so that

$$\left.\begin{aligned}
\beta_1 &= \frac{1}{R_1^{\vartriangle} A^i \rho^r c^r u} \\
\beta_2 &= 0 \\
\beta_{12} &= \frac{1}{R_{12}^{\vartriangle} A^i \rho^r c^r u} \\
\beta &= -\tfrac{\beta_1}{2} \\
\gamma &= \sqrt{\tfrac{\beta_1^2}{4} + \beta_{12}\beta_1} \\
\delta &= \tfrac{1}{\gamma}\big(\beta_{12} + \tfrac{\beta_1}{2}\big)
\end{aligned}\right\} \tag{M.101}$$

Hence, (M.82) simplifies

$$\begin{aligned}
f_1(z) &= e^{\beta z}[\cosh(\gamma z) - \delta \sinh(\gamma z)] \\
f_2(z) &= e^{\beta z}\frac{\beta_{12}}{\gamma} \sinh(\gamma z) \\
f_3(z) &= e^{\beta z}[\cosh(\gamma z) + \delta \sinh(\gamma z)] \\
f_4(z) &= e^{\beta z}\big[\beta_1 \cosh(\gamma z) - \delta\beta_1 \sinh(\gamma z)\big] \\
f_5(z) &= e^{\beta z}\frac{\beta_1\beta_{12}}{\gamma} \sinh(\gamma z)
\end{aligned} \tag{M.102}$$

The outlet temperature $T_o(t)$ is determined by

$$T_o(t) = T_i(t)\frac{f_1(L) + f_2(L)}{f_3(L) - f_2(L)} + \int_0^L \frac{T_s(\xi, t)[f_4(L-\xi) + f_5(L-\xi)]}{f_3(L) - f_2(L)} d\xi \tag{M.103}$$

and the temperature distributions T_{i1} and T_{o1} are obtained from the integral expressions[2]

$$\begin{aligned}
T_{i1}(z,t) &= T_i(t) f_1(z) + T_o(t) f_2(z) + \int_0^z T_s(\xi, t) f_4(z-\xi) d\xi \\
T_{o1}(z,t) &= -T_i(t) f_2(z) + T_o(t) f_3(z) - \int_0^z T_s(\xi, t) f_5(z-\xi) d\xi
\end{aligned} \tag{M.104}$$

[2] The integrals $f_4(z-\xi)$ and $f_5(z-\xi)$ result for the CXA configuration

$$\begin{aligned}
F_4(z,a,b) &= \textstyle\int_a^b f_4(z-\xi) d\xi \\
&= \tfrac{\beta_1}{\gamma^2-\beta^2} \exp(\beta(z-\xi))|_a^b \Big[(\gamma\delta + \beta)\cosh(\gamma(z-\xi))|_a^b - (\gamma + \delta\beta)\sinh(\gamma(z-\xi))|_a^b\Big] \\
F_5(z,a,b) &= \textstyle\int_a^b f_5(z-\xi) d\xi = \tfrac{\beta_1\beta_{12}}{\gamma^2-\beta^2} \exp(\beta(z-\xi))|_a^b \Big[\tfrac{\beta}{\gamma}\sinh(\gamma(z-\xi))|_a^b - \cosh(\gamma(z-\xi))|_a^b\Big]
\end{aligned}$$

With the equilibrium condition (Fig. M.3)

$$\frac{T_{g1} - T_s}{R_{gs}^{\mathrm{CXA}}} + \frac{T_{g1} - T_{i1}}{R_{fig}^{\mathrm{CXA}}} = 0 \tag{M.105}$$

the temperature distribution for the grout zone $T_{g1}(z,t)$ yields

$$T_{g1}(z,t) = \frac{R_{fig}^{\mathrm{CXA}}}{R_{fig}^{\mathrm{CXA}} + R_{gs}^{\mathrm{CXA}}}\big[T_s(z,t) - T_{i1}(z,t)\big] + T_{i1}(z,t) \tag{M.106}$$

The thermal resistances $R_1^{\vartriangle}$ and $R_{12}^{\vartriangle}$ are given by

$$\begin{aligned} R_1^{\vartriangle} &= R_{fig}^{\mathrm{CXA}} + R_{gs}^{\mathrm{CXA}} \\ R_{12}^{\vartriangle} &= R_{ff}^{\mathrm{CXA}} \end{aligned} \tag{M.107}$$

M.4.5 Solution for CXC Configuration

For coaxial BHE pipes with centered inlet there is

$$R_1^{\vartriangle} = \infty \tag{M.108}$$

so that

$$\left.\begin{aligned} \beta_1 &= 0 \\ \beta_2 &= \frac{1}{R_2^{\vartriangle} A^i \rho^r c^r u} \\ \beta_{12} &= \frac{1}{R_{12}^{\vartriangle} A^i \rho^r c^r u} \\ \beta &= \tfrac{\beta_2}{2} \\ \gamma &= \sqrt{\tfrac{\beta_1^2}{4} + \beta_{12}\beta_2} \\ \delta &= \tfrac{1}{\gamma}\left(\beta_{12} + \tfrac{\beta_2}{2}\right) \end{aligned}\right\} \tag{M.109}$$

Hence, (M.82) simplifies

$$\begin{aligned} f_1(z) &= e^{\beta z}[\cosh(\gamma z) - \delta \sinh(\gamma z)] \\ f_2(z) &= e^{\beta z}\frac{\beta_{12}}{\gamma}\sinh(\gamma z) \\ f_3(z) &= e^{\beta z}[\cosh(\gamma z) + \delta \sinh(\gamma z)] \end{aligned} \tag{M.110}$$

$$
\begin{aligned}
f_4(z) &= -e^{\beta z}\frac{\beta_2\beta_{12}}{\gamma}\sinh(\gamma z) \\
f_5(z) &= e^{\beta z}\big[\beta_2\cosh(\gamma z) + \delta\beta_2\sinh(\gamma z)\big]
\end{aligned}
$$

The outlet temperature $T_o(t)$ is determined by

$$
T_o(t) = T_i(t)\frac{f_1(L) + f_2(L)}{f_3(L) - f_2(L)} + \int_0^L \frac{T_s(\xi,t)[f_4(L-\xi) + f_5(L-\xi)]}{f_3(L) - f_2(L)}d\xi \tag{M.111}
$$

and the temperature distributions T_{i1} and T_{o1} are obtained from the integral expressions[3]

$$
\begin{aligned}
T_{i1}(z,t) &= T_i(t)f_1(z) + T_o(t)f_2(z) + \int_0^z T_s(\xi,t)f_4(z-\xi)d\xi \\
T_{o1}(z,t) &= -T_i(t)f_2(z) + T_o(t)f_3(z) - \int_0^z T_s(\xi,t)f_5(z-\xi)d\xi
\end{aligned} \tag{M.112}
$$

With the equilibrium condition (Fig. M.4)

$$
\frac{T_{g1} - T_s}{R_{gs}^{\text{CXC}}} + \frac{T_{g1} - T_{o1}}{R_{fog}^{\text{CXC}}} = 0 \tag{M.113}
$$

the temperature distribution for the grout zone $T_{g1}(z,t)$ yields

$$
T_{g1}(z,t) = \frac{R_{fog}^{\text{CXC}}}{R_{fog}^{\text{CXC}} + R_{gs}^{\text{CXC}}}\big[T_s(z,t) - T_{o1}(z,t)\big] + T_{o1}(z,t) \tag{M.114}
$$

The thermal resistances $R_1^{\vartriangle}$ and $R_{12}^{\vartriangle}$ are given by

$$
\begin{aligned}
R_2^{\vartriangle} &= R_{fog}^{\text{CXC}} + R_{gs}^{\text{CXC}} \\
R_{12}^{\vartriangle} &= R_{ff}^{\text{CXC}}
\end{aligned} \tag{M.115}
$$

[3] The integrals $f_4(z-\xi)$ and $f_5(z-\xi)$ result for the CXC configuration

$$
\begin{aligned}
F_4(z,a,b) &= \int_a^b f_4(z-\xi)d\xi = \tfrac{\beta_2\beta_{12}}{\beta^2-\gamma^2}\exp(\beta(z-\xi))|_a^b\Big[\tfrac{\beta}{\gamma}\sinh(\gamma(z-\xi))|_a^b - \cosh(\gamma(z-\xi))|_a^b\Big] \\
F_5(z,a,b) &= \int_a^b f_5(z-\xi)d\xi \\
&= \tfrac{\beta_2}{\beta^2-\gamma^2}\exp(\beta(z-\xi))|_a^b\Big[(\beta-\gamma\delta)\cosh(\gamma(z-\xi))|_a^b + (\delta\beta-\gamma)\sinh(\gamma(z-\xi))|_a^b\Big]
\end{aligned}
$$

M.4.6 Resulting Exchange Terms

For solving the coupled matrix system (13.37) for the BHE temperature $\boldsymbol{T}^{\pi}$ and soil temperature $\boldsymbol{T}^{s}$ the following exchange terms have to be known: $-\hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^{s})$, $-\hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^{\pi})$ and $-\boldsymbol{R}^{\pi}$. From the preceding sections we find:

$$-\hat{\boldsymbol{R}}^{\pi s}(\boldsymbol{T}^{s}) = \begin{cases} \begin{pmatrix} T_{i1}(z_i,t) \\ T_{i2}(z_i,t) \\ T_{o1}(z_i,t) \\ T_{o2}(z_i,t) \\ T_{g1}(z_i,t) \\ T_{g2}(z_i,t) \\ T_{g3}(z_i,t) \\ T_{g4}(z_i,t) \end{pmatrix} & \text{2U configuration} \\ \begin{pmatrix} T_{i1}(z_i,t) \\ T_{o1}(z_i,t) \\ T_{g1}(z_i,t) \\ T_{g2}(z_i,t) \end{pmatrix} & \text{1U configuration} \\ \begin{pmatrix} T_{i1}(z_i,t) \\ T_{o1}(z_i,t) \\ T_{g1}(z_i,t) \end{pmatrix} & \text{CXA and CXC configurations} \end{cases} \quad (1 \le i \le N_{\text{BHE}}) \tag{M.116}$$

where z_i corresponds to the z–coordinate at the BHE nodal point i, the pipe and grout temperatures are given by the analytical expressions (M.90), (M.94), (M.96), (M.104), (M.106), (M.112), and (M.114), which are dependent on the unknown soil temperature $T_s(z,t)$. Furthermore, it is

$$-\hat{\boldsymbol{R}}^{s\pi}(\boldsymbol{T}^{\pi}) = \int_z \left(\frac{\boldsymbol{T}_{i1}}{R_1^{\Delta}} + \frac{\boldsymbol{T}_{o1}}{R_2^{\Delta}} \right) dz \tag{M.117}$$

and

$$-\boldsymbol{R}^{\pi} = \delta \int_z \left(\frac{1}{R_1^{\Delta}} + \frac{1}{R_2^{\Delta}} \right) dz \tag{M.118}$$

with

$$\boldsymbol{T}_{i1} = \begin{pmatrix} T_{i1}(z_1,t) \\ T_{i1}(z_2,t) \\ \vdots \\ T_{i1}(z_{N_{\text{BHE}}},t) \end{pmatrix}, \quad \boldsymbol{T}_{o1} = \begin{pmatrix} T_{o1}(z_1,t) \\ T_{o1}(z_2,t) \\ \vdots \\ T_{o1}(z_{N_{\text{BHE}}},t) \end{pmatrix} \tag{M.119}$$

where $R_1^\vartriangle$ and $R_2^\vartriangle$ are given by (M.86), (M.98), (M.99), (M.100), (M.107), (M.108), and (M.115) for the corresponding BHE configurations.

M.5 Numerical Solution of the Local Problem

In this section, the BHE equations are solved by the numerical method which is based on the ideas given by Al-Khoury et al. [8], Al-Khoury and Bonnier [7], and Al-Khoury [6]. This method is applied to the different BHE configurations by using the extended TRCM [29]. The governing BHE heat transport equations are discretized by 1D finite element *discrete feature elements* (DFE's), cf. Chap. 14, which are coupled to the global heat transport equation for the soil (porous medium) via heat transfer relations. In comparison to the analytical method (Sect. M.4) the numerical method is more general and applicable both to short-term and long-term analyses, however, can be less efficient and stable in particular for long-term simulations.

M.5.1 Basic BHE Equations of Heat Transport in Pipes and Grout Zones

The BHE represents a closed pipe system, where a refrigerant fluid is circulating with a given velocity $\boldsymbol{u}$. The 2U configuration consists of 8 borehole components (2 pipes-in, 2 pipes-out and 4 grout zones, Fig. M.1), the 1U configuration consists of 4 borehole components (1 pipe-in, 1 pipe-out and 2 grout zones, Fig. M.2) and each of the CXA and CXC configurations consists of 3 borehole components (1 pipe-in, 1 pipe-out and 1 grout zone, Figs. M.3 and M.4), which are schematized by the corresponding number of 1D DFE's as shown in Fig. M.6. Their 1D heat transport equations are given as follows for the pipe(s)

$$\begin{aligned} \rho^r c^r \frac{\partial T_k}{\partial t} + \rho^r c^r \boldsymbol{u} \cdot \nabla T_k - \nabla \cdot (\boldsymbol{\Lambda}^r \cdot \nabla T_k) &= H_k \qquad \text{in} \quad \Omega_k \\ \text{with Cauchy-type BC:} \quad -(\boldsymbol{\Lambda}^r \cdot \nabla T_k) \cdot \boldsymbol{n} &= q_{nT_k} \qquad \text{on} \quad \Gamma_k \\ \text{for} \quad k = i1, o1, (i2, o2) & \end{aligned} \tag{M.120}$$

and for the grout zone(s)

$$\begin{aligned} \rho^g c^g \frac{\partial T_k}{\partial t} - \nabla \cdot (\Lambda^g \nabla T_k) &= H_k \qquad \text{in} \quad \Omega_k \\ \text{with Cauchy-type BC:} \quad -(\Lambda^g \nabla T_k) \cdot \boldsymbol{n} &= q_{nT_k} \qquad \text{on} \quad \Gamma_k \\ \text{for} \quad k = g1, (g2, (g3, g4)) & \end{aligned} \tag{M.121}$$

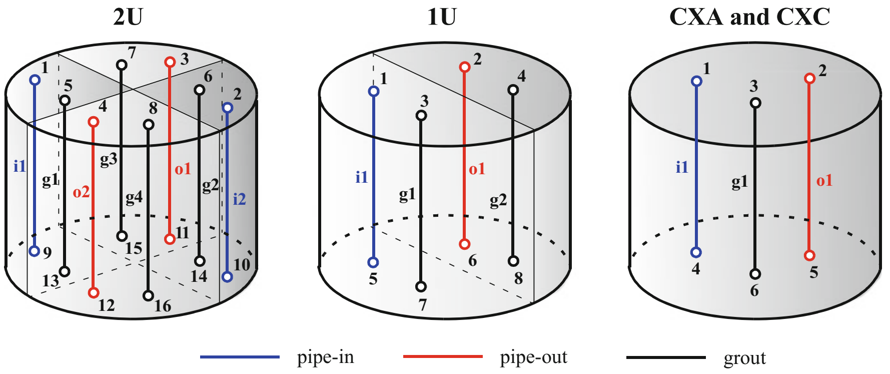


Fig. M.6 BHE pipe(s)-in, pipe(s)-out and grout zone components for 2U, 1U, CXA and CXC configurations discretized by 1D DFE's. Local node numbering is used for the vertical 1D representations of the components

which are considered impervious, where

$$\boldsymbol{\Lambda}^r = (\Lambda^r + \rho^r c^r \beta_L \|\boldsymbol{u}\|)\boldsymbol{\delta} \tag{M.122}$$

is the hydrodynamic thermodispersion for the refrigerant, $\nabla = \partial/\partial z$ is defined here for the vertical 1D line direction z along the BHE 'well' axis and the boundary heat fluxes q_{nT_k} are summarized in Table M.2 for the related BHE components k. Note that $\Phi_{ff}^{\mathrm{CXA}} \neq \Phi_{ff}^{\mathrm{CXC}}$ due to the different pipe radii for pipe-in and pipe-out in a coaxial pipe installation.

M.5.2 Finite Element Discretization of the BHE Equations

The BHE Eqs. (M.120) and (M.121) are discretized by finite elements. Introducing the spatial weighting function w the following weak statements result for the pipe(s) components

$$\begin{aligned}&\int_{\Omega_k}\Big[w\rho^r c^r\Big(\frac{\partial T_k}{\partial t} + \boldsymbol{u}\cdot\nabla T_k\Big) + \nabla w\cdot(\boldsymbol{\Lambda}^r\cdot\nabla T_k)\Big]d\Omega = \\ &-\int_{\Gamma_k} w q_{nT_k} d\Gamma + \int_{\Omega_k} wH_k d\Omega \quad \text{for} \quad k = i1, o1, (i2, o2)\end{aligned} \tag{M.123}$$

and for the grout zone(s) components of the BHE

$$\begin{aligned}&\int_{\Omega_k}\Big[w\rho^g c^g \frac{\partial T_k}{\partial t} + \nabla w\cdot(\Lambda^g \nabla T_k)\Big]d\Omega = \\ &-\int_{\Gamma_k} w q_{nT_k} d\Gamma + \int_{\Omega_k} wH_k d\Omega \quad \text{for} \quad k = g1, (g2, (g3, g4))\end{aligned} \tag{M.124}$$

Table M.2 Boundary heat fluxes q_{nT_k} for the BHE components k

	$q_{nT_k} =$			
k	2U	1U	CXA	CXC
i1	$-\Phi_{fig}^{2U}(T_{g1} - T_{i1})$	$-\Phi_{fig}^{1U}(T_{g1} - T_{i1})$	$-\Phi_{fig}^{CXA}(T_{g1} - T_{i1}) - \Phi_{ff}^{CXA}(T_{o1} - T_{i1})$	$-\Phi_{ff}^{CXC}(T_{o1} - T_{i1})$
i2	$-\Phi_{fig}^{2U}(T_{g2} - T_{i2})$	–	–	–
o1	$-\Phi_{fog}^{2U}(T_{g3} - T_{o1})$	$-\Phi_{fog}^{1U}(T_{g2} - T_{o1})$	$-\Phi_{ff}^{CXA}(T_{i1} - T_{o1})$	$-\Phi_{fog}^{CXC}(T_{g1} - T_{o1}) - \Phi_{ff}^{CXC}(T_{i1} - T_{o1})$
o2	$-\Phi_{fog}^{2U}(T_{g4} - T_{o2})$	–	–	–
g1	$-\Phi_{gs}^{2U}(T_s - T_{g1}) - \Phi_{fig}^{2U}(T_{i1} - T_{g1}) - \Phi_{gg2}^{2U}(T_{g2} - T_{g1}) - \Phi_{gg1}^{2U}(T_{g3} - T_{g1}) - \Phi_{gg1}^{2U}(T_{g4} - T_{g1})$	$-\Phi_{gs}^{1U}(T_s - T_{g1}) - \Phi_{fig}^{1U}(T_{i1} - T_{g1}) - \Phi_{gg}^{1U}(T_{g2} - T_{g1})$	$-\Phi_{gs}^{CXA}(T_s - T_{g1}) - \Phi_{fig}^{CXA}(T_{i1} - T_{g1})$	$-\Phi_{gs}^{CXC}(T_s - T_{g1}) - \Phi_{fog}^{CXC}(T_{o1} - T_{g1})$
g2	$-\Phi_{gs}^{2U}(T_s - T_{g2}) - \Phi_{fig}^{2U}(T_{i2} - T_{g2}) - \Phi_{gg2}^{2U}(T_{g1} - T_{g2}) - \Phi_{gg1}^{2U}(T_{g3} - T_{g2}) - \Phi_{gg1}^{2U}(T_{g4} - T_{g2})$	$-\Phi_{gs}^{1U}(T_s - T_{g2}) - \Phi_{fog}^{1U}(T_{o1} - T_{g2}) - \Phi_{gg}^{1U}(T_{g1} - T_{g2})$	–	–
g3	$-\Phi_{gs}^{2U}(T_s - T_{g3}) - \Phi_{fog}^{2U}(T_{o1} - T_{g3}) - \Phi_{gg2}^{2U}(T_{g4} - T_{g3}) - \Phi_{gg1}^{2U}(T_{g1} - T_{g3}) - \Phi_{gg1}^{2U}(T_{g2} - T_{g3})$	–	–	–
g4	$-\Phi_{gs}^{2U}(T_s - T_{g4}) - \Phi_{fog}^{2U}(T_{o2} - T_{g4}) - \Phi_{gg2}^{2U}(T_{g3} - T_{g4}) - \Phi_{gg1}^{2U}(T_{g1} - T_{g4}) - \Phi_{gg1}^{2U}(T_{g2} - T_{g4})$	–	–	–

where the boundary heat exchange q_{nT_k} is specified in Table M.2. Using GFEM (M.123) and (M.124) lead to the following matrix system

$$\boldsymbol{P}^{\pi} \cdot \dot{\boldsymbol{T}}^{\pi} + \boldsymbol{L}^{\pi} \cdot \boldsymbol{T}^{\pi} = \boldsymbol{W}^{\pi} - \boldsymbol{R}^{\pi s} \cdot \boldsymbol{T}^{s} \tag{M.125}$$

with

$$\boldsymbol{P}^{\pi} = \begin{cases} \begin{pmatrix} \boldsymbol{P}_{i1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{i2} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{o1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{o2} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{g1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{g2} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{g3} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{g4} \end{pmatrix} & \text{2U} \\ \begin{pmatrix} \boldsymbol{P}_{i1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{o1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{g1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{g2} \end{pmatrix} & \text{1U} \\ \begin{pmatrix} \boldsymbol{P}_{i1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_{o1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{P}_{g1} \end{pmatrix} & \text{CXA or CXC} \end{cases} \tag{M.126}$$

$$\boldsymbol{L}^{\pi} = \begin{cases} \begin{pmatrix} \boldsymbol{K}_{i1} & \boldsymbol{0} & -\boldsymbol{R}_{io} & \boldsymbol{0} & -\boldsymbol{R}_{i1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K}_{i2} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{R}_{i2} & \boldsymbol{0} & \boldsymbol{0} \\ -\boldsymbol{R}_{io} & \boldsymbol{0} & \boldsymbol{K}_{o1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{R}_{o1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{K}_{o2} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{R}_{o2} \\ -\boldsymbol{R}_{i1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{K}_{ig} & -\boldsymbol{R}_{g2} & -\boldsymbol{R}_{g1} & -\boldsymbol{R}_{g1} \\ \boldsymbol{0} & -\boldsymbol{R}_{i2} & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{R}_{g2} & \boldsymbol{K}_{ig} & -\boldsymbol{R}_{g1} & -\boldsymbol{R}_{g1} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{R}_{o1} & \boldsymbol{0} & -\boldsymbol{R}_{g1} & -\boldsymbol{R}_{g1} & \boldsymbol{K}_{og} & -\boldsymbol{R}_{g2} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{R}_{o2} & -\boldsymbol{R}_{g1} & -\boldsymbol{R}_{g1} & -\boldsymbol{R}_{g2} & \boldsymbol{K}_{og} \end{pmatrix} & \text{2U} \\ \begin{pmatrix} \boldsymbol{K}_{i1} & -\boldsymbol{R}_{io} & -\boldsymbol{R}_{i1} & \boldsymbol{0} \\ -\boldsymbol{R}_{io} & \boldsymbol{K}_{o1} & \boldsymbol{0} & -\boldsymbol{R}_{o1} \\ -\boldsymbol{R}_{i1} & \boldsymbol{0} & \boldsymbol{K}_{ig} & -\boldsymbol{R}_{g1} \\ \boldsymbol{0} & -\boldsymbol{R}_{o1} & -\boldsymbol{R}_{g1} & \boldsymbol{K}_{og} \end{pmatrix} & \text{1U} \\ \begin{pmatrix} \boldsymbol{K}_{i1} & -\boldsymbol{R}_{io} & -\boldsymbol{R}_{i1} \\ -\boldsymbol{R}_{io} & \boldsymbol{K}_{o1} & \boldsymbol{0} \\ -\boldsymbol{R}_{i1} & \boldsymbol{0} & \boldsymbol{K}_{ig} \end{pmatrix} & \text{CXA} \\ \begin{pmatrix} \boldsymbol{K}_{i1} & -\boldsymbol{R}_{io} & \boldsymbol{0} \\ -\boldsymbol{R}_{io} & \boldsymbol{K}_{o1} & -\boldsymbol{R}_{o1} \\ \boldsymbol{0} & -\boldsymbol{R}_{o1} & \boldsymbol{K}_{og} \end{pmatrix} & \text{CXC} \end{cases} \tag{M.127}$$

$$
\boldsymbol{R}^{\pi s} = \left\{ \begin{array}{l}
\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ -\boldsymbol{R}_s \\ -\boldsymbol{R}_s \\ -\boldsymbol{R}_s \\ -\boldsymbol{R}_s \end{pmatrix} \\
\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ -\boldsymbol{R}_s \\ -\boldsymbol{R}_s \end{pmatrix} \\
\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ -\boldsymbol{R}_s \end{pmatrix}
\end{array} \right. , \quad
\boldsymbol{W}^{\pi} = \left\{ \begin{array}{l}
\begin{pmatrix} \boldsymbol{W}_{i1} \\ \boldsymbol{W}_{i2} \\ \boldsymbol{W}_{o1} \\ \boldsymbol{W}_{o2} \\ \boldsymbol{W}_{g1} \\ \boldsymbol{W}_{g2} \\ \boldsymbol{W}_{g3} \\ \boldsymbol{W}_{g4} \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{W}_{i1} \\ \boldsymbol{W}_{o1} \\ \boldsymbol{W}_{g1} \\ \boldsymbol{W}_{g2} \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{W}_{i1} \\ \boldsymbol{W}_{o1} \\ \boldsymbol{W}_{g1} \end{pmatrix}
\end{array} \right. , \quad
\boldsymbol{T}^{\pi} = \left\{ \begin{array}{ll}
\begin{pmatrix} \boldsymbol{T}_{i1} \\ \boldsymbol{T}_{i2} \\ \boldsymbol{T}_{o1} \\ \boldsymbol{T}_{o2} \\ \boldsymbol{T}_{g1} \\ \boldsymbol{T}_{g2} \\ \boldsymbol{T}_{g3} \\ \boldsymbol{T}_{g4} \end{pmatrix} & \text{2U} \\
\begin{pmatrix} \boldsymbol{T}_{i1} \\ \boldsymbol{T}_{o1} \\ \boldsymbol{T}_{g1} \\ \boldsymbol{T}_{g2} \end{pmatrix} & \text{1U} \\
\begin{pmatrix} \boldsymbol{T}_{i1} \\ \boldsymbol{T}_{o1} \\ \boldsymbol{T}_{g1} \end{pmatrix} & \text{CXA or CXC}
\end{array} \right. \tag{M.128}
$$

and similarly for $\dot{\boldsymbol{T}}^{\pi}$, where

$$
\boldsymbol{P}_k = \begin{cases}
\sum_e \int_{\Omega_k^e} \rho^r c^r N_i N_j \, d\Omega^e & \text{for} \quad k = i1, o1, (i2, o2) \\
\sum_e \int_{\Omega_k^e} \rho^g c^g N_i N_j \, d\Omega^e & \text{for} \quad k = g1, (g2, (g3, g4))
\end{cases} \tag{M.129}
$$

$$
\begin{aligned}
\boldsymbol{K}_{i1} &= \boldsymbol{C}_{i1} + \boldsymbol{R}_{i1} + \boldsymbol{R}_{io} \\
\boldsymbol{K}_{i2} &= \boldsymbol{C}_{i2} + \boldsymbol{R}_{i2} \\
\boldsymbol{K}_{o1} &= \boldsymbol{C}_{o1} + \boldsymbol{R}_{o1} + \boldsymbol{R}_{io} \\
\boldsymbol{K}_{o2} &= \boldsymbol{C}_{o2} + \boldsymbol{R}_{o2} \\
\boldsymbol{K}_{ig} &= \boldsymbol{G}_{ig} + \boldsymbol{R}_{i1} + 2\boldsymbol{R}_{g1} + \boldsymbol{R}_{g2} + \boldsymbol{R}_s \\
\boldsymbol{K}_{og} &= \boldsymbol{G}_{og} + \boldsymbol{R}_{o1} + 2\boldsymbol{R}_{g1} + \boldsymbol{R}_{g2} + \boldsymbol{R}_s
\end{aligned} \tag{M.130}
$$

$$
\begin{aligned}
\boldsymbol{C}_k &= \sum_e \int_{\Omega_k^e} \big(N_i \rho^r c^r \boldsymbol{u} \cdot \nabla N_j \nabla N_i \cdot (\boldsymbol{\Lambda} \cdot \nabla N_j) \big) d\Omega^e && \text{for} \quad k = i1, o1, (i2, o2) \\
\boldsymbol{G}_{ig} = \boldsymbol{G}_{og} &= \sum_e \int_{\Omega_k^e} \nabla N_i \cdot (\Lambda^g \nabla N_j) d\Omega^e && \text{for} \quad k = (g1, g2, g3, g4) \\
\boldsymbol{W}_k &= \sum_e \int_{\Omega_k^e} N_i H_k d\Omega^e && \text{for} \quad \forall k
\end{aligned} \tag{M.131}
$$

and the boundary heat exchange matrices as listed in Table M.3, in which $(i, j = 1, \ldots, N_{\mathrm{BHE}})$ and e runs over the associated numbers of 1D DFE's. The symbols $\Omega^e_{i1,i2}$, $\Gamma^e_{i1,i2}$ denote the domain and surface of pipe(s)-in, $\Omega^e_{o1,o2}$, $\Gamma^e_{o1,o2}$ for pipe(s)-out and Ω^e_{gi}, Γ^e_{gi} $(i = 1, \ldots, G)$ for the grout zones (Γ^e_{g1-g4} for all grout zones) of finite element e (G is the number of grout zones: 4 for 2U, 2 for 1U and 1 for CXA and CXC configurations). Furthermore, the following heat transfer matrices are defined:

$$\boldsymbol{R}^{s\pi} = \boldsymbol{R}^{\pi s^T} = \begin{cases} (\mathbf{0}\ \mathbf{0}\ \mathbf{0}\ \mathbf{0}\ -\boldsymbol{R}_s\ -\boldsymbol{R}_s\ -\boldsymbol{R}_s\ -\boldsymbol{R}_s) & \text{2U} \\ (\mathbf{0}\ \mathbf{0}\ -\boldsymbol{R}_s\ -\boldsymbol{R}_s) & \text{1U} \\ (\mathbf{0}\ \mathbf{0}\ -\boldsymbol{R}_s) & \text{CXA or CXC} \end{cases} \tag{M.132}$$

and

$$\boldsymbol{R}^{\pi} = -G\boldsymbol{R}_s \tag{M.133}$$

which appear in the coupled matrix system (13.44).

Analytical integration of matrices (M.129)–(M.131) and those of Table M.3 is performed for linear 1D elements in accordance with Sect. H.1 of Appendix H. In doing so, the volume and surface elements are

$$\begin{aligned} d\Omega^e &= A dz = A|\boldsymbol{J}^e| d\xi = \frac{A\Delta z^e}{2} d\xi \\ d\Gamma^e &= S dz = S|\boldsymbol{J}^e| d\xi = \frac{S\Delta z^e}{2} d\xi \end{aligned} \tag{M.134}$$

where $|\boldsymbol{J}^e| = \Delta z^e/2$ is the Jacobian (H.4), Δz^e is the length of element e, ξ is the local coordinate, S is the specific exchange surface given in Table M.1 and A is a cross-sectional area given in Table M.4 for the inner pipes and grout zones. The used linear 1D element e consists of two nodes with each DOF temperature variables for the BHE components $i1, i2, o1, o2, g1, g2, g3, g4$ and with only one temperature variable for the soil s as depicted in Fig. M.7.

In using these relationships we find the matrices for element e as follows:

$$\begin{aligned} \boldsymbol{P}^e_{i1,i2} &= \frac{A^i_i \rho^r c^r \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\ \boldsymbol{P}^e_{o1,o2} &= \frac{A^i_o \rho^r c^r \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\ \boldsymbol{P}^e_{g1,g2} &= \frac{A^i_g \rho^g c^g \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\ \boldsymbol{P}^e_{g3,g4} &= \frac{A^o_g \rho^g c^g \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \end{aligned} \tag{M.135}$$

Table M.3 Boundary heat exchange matrices

	2U	1U	CXA	CXC
$\boldsymbol{R}_{i1}$ $\boldsymbol{R}_{i2}$	$\sum_e \int_{\Gamma^e_{i1,i2}} \Phi^{2\mathrm{U}}_{fig} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{i1}} \Phi^{1\mathrm{U}}_{fig} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{i1}} \Phi^{\mathrm{CXA}}_{fig} N_i N_j \, d\Gamma^e$	$\mathbf{0}$
$\boldsymbol{R}_{o1}$ $\boldsymbol{R}_{o2}$	$\sum_e \int_{\Gamma^e_{o1,o2}} \Phi^{2\mathrm{U}}_{fog} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{o1}} \Phi^{1\mathrm{U}}_{fog} N_i N_j \, d\Gamma^e$	$\mathbf{0}$	$\sum_e \int_{\Gamma^e_{o1}} \Phi^{\mathrm{CXC}}_{fog} N_i N_j \, d\Gamma^e$
$\boldsymbol{R}_{io}$	$\mathbf{0}$	$\mathbf{0}$	$\sum_e \int_{\Gamma^e_{o1}} \Phi^{\mathrm{CXA}}_{ff} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{i1}} \Phi^{\mathrm{CXC}}_{ff} N_i N_j \, d\Gamma^e$
$\boldsymbol{R}_{g1}$	$\sum_e \int_{\Gamma^e_{g1-g4}} \Phi^{2\mathrm{U}}_{gg1} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{g1,g2}} \Phi^{1\mathrm{U}}_{gg} N_i N_j \, d\Gamma^e$	$\mathbf{0}$	$\mathbf{0}$
$\boldsymbol{R}_{g2}$	$\sum_e \int_{\Gamma^e_{g1-g4}} \Phi^{2\mathrm{U}}_{gg2} N_i N_j \, d\Gamma^e$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
$\boldsymbol{R}_{s}$	$\sum_e \int_{\Gamma^e_{g1-g4}} \Phi^{2\mathrm{U}}_{gs} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{g1,g2}} \Phi^{1\mathrm{U}}_{gs} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{g1}} \Phi^{\mathrm{CXA}}_{gs} N_i N_j \, d\Gamma^e$	$\sum_e \int_{\Gamma^e_{g1}} \Phi^{\mathrm{CXC}}_{gs} N_i N_j \, d\Gamma^e$

Table M.4 Cross-sectional areas A for the BHE configurations according to Figs. M.1–M.4

A	2U[a]	1U	CXA	CXC
A_i^i	$\pi(r_i^i)^2$	$\pi(r_i^i)^2$	$\pi[(r_i^i)^2-(r_o^o)^2]$	$\pi(r_i^i)^2$
A_o^i	$\pi(r_o^i)^2$	$\pi(r_o^i)^2$	$\pi(r_o^i)^2$	$\pi[(r_o^i)^2-(r_i^o)^2]$
A_g^i	$\pi\big[\frac{1}{4}\frac{D^2}{4}-(r_i^o)^2\big]$	$\pi\big[\frac{1}{2}\frac{D^2}{4}-(r_i^o)^2\big]$	$\pi\big[\frac{D^2}{4}-(r_i^o)^2\big]$	–
A_g^o	$\pi\big[\frac{1}{4}\frac{D^2}{4}-(r_o^o)^2\big]$	$\pi\big[\frac{1}{2}\frac{D^2}{4}-(r_o^o)^2\big]$	–	$\pi\big[\frac{D^2}{4}-(r_o^o)^2\big]$

[a] In case of 2U exchangers it is assumed that the radii for the two pipes-in and two pipes-out are equal, i.e., it is defined: $r_i^i=r_{i1}^i=r_{i2}^i$, $r_i^o=r_{i1}^o=r_{i2}^o$, $r_o^i=r_{o1}^i=r_{o2}^i$ and $r_o^o=r_{o1}^o=r_{o2}^o$

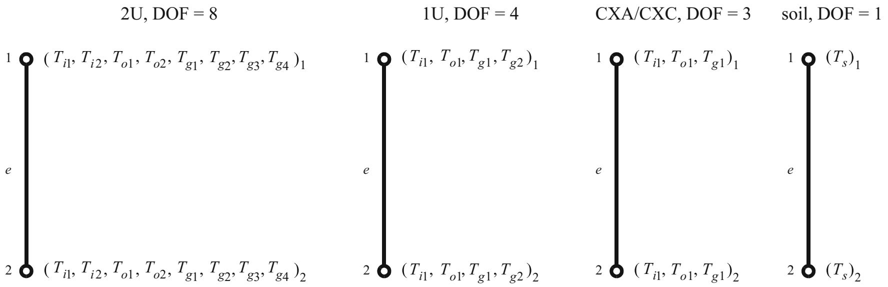

2U, DOF = 8
1U, DOF = 4
CXA/CXC, DOF = 3
soil, DOF = 1
1
$(T_{i1}, T_{i2}, T_{o1}, T_{o2}, T_{g1}, T_{g2}, T_{g3}, T_{g4})_1$
e
2
$(T_{i1}, T_{i2}, T_{o1}, T_{o2}, T_{g1}, T_{g2}, T_{g3}, T_{g4})_2$
1
$(T_{i1}, T_{o1}, T_{g1}, T_{g2})_1$
e
2
$(T_{i1}, T_{o1}, T_{g1}, T_{g2})_2$
1
$(T_{i1}, T_{o1}, T_{g1})_1$
e
2
$(T_{i1}, T_{o1}, T_{g1})_2$
1
$(T_s)_1$
e
2
$(T_s)_2$

Fig. M.7 1D element e used for BHE components of 2U, 1U, CXA or CXC configuration and for soil s

$$\begin{aligned}
\boldsymbol{C}^e_{i1,i2} &= \frac{A_i^i\rho^r c^r u}{2}\begin{pmatrix}-1 & 1\\ -1 & 1\end{pmatrix} + \frac{A_i^i\|\boldsymbol{\Lambda}^r\|}{\Delta z^e}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\\
\boldsymbol{C}^e_{o1,o2} &= \frac{A_o^i\rho^r c^r(-u)}{2}\begin{pmatrix}-1 & 1\\ -1 & 1\end{pmatrix} + \frac{A_o^i\|\boldsymbol{\Lambda}^r\|}{\Delta z^e}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}
\end{aligned} \tag{M.136}$$

$$\begin{aligned}
\boldsymbol{G}^e_{ig} &= \frac{A_g^i\Lambda^g}{\Delta z^e}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\\
\boldsymbol{G}^e_{og} &= \frac{A_g^o\Lambda^g}{\Delta z^e}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}
\end{aligned} \tag{M.137}$$

$$\begin{aligned}
\boldsymbol{W}^e_{i1,i2} &= \frac{A_i^i H_{i1,i2}\,\Delta z^e}{2}\begin{pmatrix}1\\ 1\end{pmatrix}\\
\boldsymbol{W}^e_{o1,o2} &= \frac{A_o^i H_{o1,o2}\,\Delta z^e}{2}\begin{pmatrix}1\\ 1\end{pmatrix}\\
\boldsymbol{W}^e_{g1,g2} &= \frac{A_g^i H_{g1,g2}\,\Delta z^e}{2}\begin{pmatrix}1\\ 1\end{pmatrix}\\
\boldsymbol{W}^e_{g3,g4} &= \frac{A_g^o H_{g3,g4}\,\Delta z^e}{2}\begin{pmatrix}1\\ 1\end{pmatrix}
\end{aligned} \tag{M.138}$$

$$
\begin{aligned}
\boldsymbol{R}^e_{i1,i2} &= \frac{\Phi_{fig}\, S_i\, \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} &= \frac{\Delta z^e}{6R_{fig}} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{R}^e_{o1,o2} &= \frac{\Phi_{fog}\, S_o\, \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} &= \frac{\Delta z^e}{6R_{fog}} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{R}^e_{io} &= \frac{\Phi_{ff}\, S_{io}\, \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} &= \frac{\Delta z^e}{6R_{ff}} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{R}^e_{g1} &= \frac{\Phi_{gg1}\, S_{g1}\, \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} &= \frac{\Delta z^e}{6R_{gg1}} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{R}^e_{g2} &= \frac{\Phi_{gg2}\, S_{g2}\, \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} &= \frac{\Delta z^e}{6R_{gg2}} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{R}^e_{s} &= \frac{\Phi_{gs}\, S_{gs}\, \Delta z^e}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} &= \frac{\Delta z^e}{6R_{gs}} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}
\end{aligned}
\tag{M.139}
$$

If the advective part in the heat transport equations of the BHE pipes becomes dominant, wiggles in the temperature solutions can occur and the spatial discretization with the standard GFEM becomes insufficient. A common technique is the SU scheme (cf. Sect. 8.14.3), which introduces a balancing diffusivity to produce stabilized wiggle-free (smooth) solutions. It is equivalent to modifying the thermal dispersion tensor (M.122) for the refrigerant of the 1D pipes according to

$$
\boldsymbol{\Lambda}^r = \big(\Lambda^r + \rho^r c^r (\beta_L + \beta_{\mathrm{num}}) \|\boldsymbol{u}\|\big)\boldsymbol{\delta}
\tag{M.140}
$$

with a numerical thermodispersivity $\beta_{\mathrm{num}} = \Delta z^e/2$ derived for linear finite elements (cf. Table 8.9).

References

1. Abarca, E., Carrera, J., Sánchez-Vila, X., Voss, C.: Quasi-horizontal circulation cells in 3D seawater intrusion. J. Hydrol. **339**(3–4), 118–129 (2007)
2. Ackerer, P., Younes, A., Mosé, R.: Modeling variable density flow and solute transport in porous medium: 1. Numerical model and verification. Transp. Porous Media **35**(3), 345–373 (1999)
3. Ackerer, P., Younes, A., Oswald, S., Kinzelbach, W.: On modeling of density driven flow. In: Stauffer, F., et al. (eds.) MODELCARE99 – Calibration and Reliability in Groundwater Modelling: Coping with Uncertainty, Zurich, 1999. IAHS Publication, No. 265, pp. 377–384. IAHS (2000)
4. ADEME: MACAOH (2001–2006): Modélisation du devenir des composés organo-chlorés aliphatiques dans les aquiféres. Technical report, French Environment and Energy Management Agency (2007). http://www2.ademe.fr/publication
5. Adler, P., Thovert, J.F.: Fractures and Fracture Networks. Kluwer Academic, Dordrecht (1999)
6. Al-Khoury, R.: Computational Modeling of Shallow Geothermal Systems. CRC/Balkema/Taylor & Francis, London (2012)
7. Al-Khoury, R., Bonnier, P.: Efficient finite element formulation for geothermal heating systems. Part II: Transient. Int. J. Numer. Methods Eng. **67**(5), 725–745 (2006)
8. Al-Khoury, R., Bonnier, P., Brinkgreve, R.: Efficient finite element formulation for geothermal heating systems. Part I: Steady state. Int. J. Numer. Methods Eng. **63**(7), 988–1013 (2005)
9. Anderson, M., Woessner, W.: Applied groundwater modeling – simulation of flow and advective transport. Academic, San Diego (1992)
10. Argyris, J., Vaz, L., Willam, K.: Higher order methods for transient diffusion analysis. Comput. Methods Appl. Mech. Eng. **12**(2), 243–278 (1977)
11. Aricò, C., Sinagra, M., Tucciarelli, T.: The MAST-edge centred lumped scheme for the flow simulation in variably saturated heterogeneous porous media. J. Comput. Phys. **231**(4), 1387–1425 (2012)
12. Aris, R.: Vectors, Tensors, and the Basis Equations of Fluid Mechanics. Dover, New York (1962)
13. Atkins, P.: Physical Chemistry, 5th edn. Oxford University Press, Oxford (1994)
14. Austin, W., Yavuzturk, C., Spitler, J.: Development of an in-situ system and analysis procedure for measuring ground thermal properties. ASHRAE Trans. **106**(1), 356–379 (2000)
15. Axelsson, O.: Iterative Solution Methods. Cambridge University Press, Cambridge (1994)
16. Babuška, I.: Reliability of computational mechanics. In: Whiteman, J. (ed.) The Mathematics of Finite Elements and Applications: Highlights 1993, pp. 25–44. Wiley, Chichester (1994)
17. Babuška, I., Banerjee, U.: Stable generalized finite element method (SGFEM). Comput. Methods Appl. Mech. Eng. **201–204**, 91–111 (2012)

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,

18. Babuška, I., Miller, A.: The post-processing approach in the finite element method – part 1: calculation of displacements, stresses and other higher derivatives of the displacements. Int. J. Numer. Methods Eng. **20**(6), 1085–1109 (1984)
19. Badon-Ghyben, W.: Nota in verband met de voorgenomen putboring nabij Amsterdam (notes on the probable results of well drilling near Amsterdam). In: Tijdschrift van het Kononklijk Instituut van Ingenieurs, vol. 9, pp. 8–22. The Hague (1888)
20. Baker, A.: Finite element method (Chapter 28). In: Johnson, R. (ed.) The Handbook of Fluid Dynamics, pp. 28:1–98. CRC/Springer, Boca Raton/Heidelberg (1998)
21. Bakhvalov, N.: On the convergence of a relaxation method with natural constraints on the elliptic operator. USSR Comput. Math. Math. Phys. **6**(5), 101–135 (1966)
22. Bakker, M., Hemker, K.: Analytical solutions for groundwater whirls in box-shaped, layered anisotropic aquifers. Adv. Water Resour. **27**(11), 1075–1086 (2004)
23. Bank, R.: PLTMG: a software package for solving elliptic partial differential equations – user's guide 11.0. Technical report, Department of Mathematics, University of California at San Diego, La Jolla (2012). http://ccom.ucsd.edu/~reb/software.html
24. Bank, R., Sherman, A., Weiser, A.: Refinement algorithms and data structure for regular local mesh refinement. In: Steplemen, R., et al. (eds.) Scientific Computing, pp. 3–17. IMACS/North Holland, Brussels (1983)
25. Banks, D.: An Introduction to Thermogeology: Ground Source Heating and Cooling. Blackwell, Oxford (2008)
26. Barenblatt, G., Entov, V., Ryzhik, V.: Theory of Fluid Flows Through Natural Rocks. Kluwer Academic, Dordrecht (1990)
27. Bathe, K.J., Khoshgoftaar, M.: Finite element free surface seepage analysis without mesh iteration. Int. J. Numer. Anal. Methods Geomech. **3**(1), 13–22 (1979)
28. Bauer, D., Heidemann, W., Diersch, H.J.: Transient 3D analysis of borehole heat exchanger modeling. Geothermics **40**(4), 250–260 (2011)
29. Bauer, D., Heidemann, W., Müller-Steinhagen, H., Diersch, H.J.: Thermal resistance and capacity models for borehole heat exchangers. Int. J. Energy Res. **35**(4), 312–320 (2011)
30. Bause, M.: Higher and lowest order mixed finite element approximation of subsurface flow problems with solutions of low regularity. Adv. Water Resour. **31**(2), 370–382 (2008)
31. Bause, M., Knabner, P.: Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods. Adv. Water Resour. **27**(6), 561–581 (2004)
32. Baxter, G., Wallace, C.: Changes in volume upon solution in water of the halogen salts of the alkali metals. J. Am. Chem. Soc. **38**(1), 70–105 (1916)
33. Bear, J.: Dynamics of Fluids in Porous Media. American Elsevier, New York (1972)
34. Bear, J.: Hydraulics of Groundwater. McGraw-Hill, New York (1979)
35. Bear, J.: Modeling flow and contaminant transport in fractured rocks. In: Bear, J., et al. (eds.) Flow and Contaminat Transport in Fractured Rock, pp. 1–37. Academic, San Diego (1993)
36. Bear, J.: Conceptual and mathematical modeling. In: Bear, J., et al. (eds.) Seawater Intrusion in Coastal Aquifers – Concepts, Methods and Practices, pp. 127–161. Kluwer Academic, Dordrecht (1999)
37. Bear, J., Bachmat, Y.: Introduction to Modeling of Transport Phenomena in Porous Media. Kluwer Academic, Dordrecht (1991)
38. Bear, J., Cheng, A.D.: Modeling Groundwater Flow and Contaminant Transport. Springer, Dordrecht (2010)
39. Bear, J., Verruijt, A.: Modeling Groundwater Flow and Pollution. D. Reidel, Dordrecht (1987)
40. Beauwens, R.: Modfied incomplete factorization strategies. In: Axelsson, O., Kolotilina, L. (eds.) Precondioned Conjugate Gradient Methods. Lecture Notes in Mathematics, vol. 1457, pp. 1–16. Springer, Berlin/Heidelberg/New York (1990)
41. Beck, J.: Convection in a box of porous material saturated with fluid. Phys. Fluids **15**(8), 1377–1383 (1972)
42. Behie, A., Vinsome, P.: Block iterative methods for fully implicit reservoir simulation. SPE J. **22**(5), 658–668 (1982)
43. Bejan, A., Kraus, A.: Handbook of Heat Transfer, 1st edn. Wiley, Hoboken (2003)

44. Belytschko, T., Lu, Y., Gu, L.: Element-free Galerkin methods. Int. J. Numer. Methods Eng. **37**(2), 229–256 (1994)
45. Benzi, M.: Preconditioning techniques for large linear systems: a survey. J. Comput. Phys. **182**(2), 418–477 (2002)
46. Bergamaschi, L., Putti, M.: Mixed finite elements and Newton-like linearization for the solution of Richard's equation. Int. J. Numer. Methods Eng. **45**(8), 1025–1046 (1999)
47. Berger, R., Howington, S.: Discrete fluxes and mass balance in finite elements. J. Hydraul. Eng. **128**(1), 87–92 (2002)
48. Bixler, N.: An improved time integrator for finite element analysis. Commun. Appl. Numer. Methods **5**(2), 69–78 (1989)
49. Boussinesq, J.: Théorie analytique de la chaleur, vol. 2. Gauthier-Villars, Paris (1903)
50. Bowyer, A.: Computing Dirichlet tesselations. Comput. J. **24**(2), 162–166 (1981)
51. Brandt, A.: Multi-level adaptive solutions to boundary-value problems. Math. Comput. **31**(138), 333–390 (1977)
52. Brandt, A.: Algebraic multigrid theory: the symmetric case. Appl. Math. Comput. **19**(1–4), 23–56 (1986)
53. Brandt, A., Fernando, H. (eds.): Double-Diffusive Convection. Geophysical Monograph, vol. 94. American Geophysical Union, Washington, DC (1995)
54. Brebbia, C., Telles, J., Wrobel, L.: Boundary Element Methods – Theory and Applications. Springer, New York (1983)
55. Brenner, S., Scott, L.: The Mathematical Theory of Finite Element Methods. Springer, Berlin (1994)
56. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, New York (1991)
57. Brooks, A., Hughes, T.: Streamlin upwind/Petrov-Galerkin formulations for convective dominated flow with particular emphasis on the incompressible Navier-Sokes equations. Comput. Methods Appl. Mech. Eng. **32**(1–3), 199–259 (1982)
58. Brooks, R., Corey, A.: Properties of porous media affecting fluid flow. J. Irrig. Drain. Div. Proc. ASCE **92**(IR2), 61–88 (1966)
59. Brown, G.: Henry Darcy and the making of a law. Water Resour. Res. **38**(7) (2002). doi:10.1029/2001WR000727
60. Bruch, J., Street, R.: Two-dimensional dispersion. J. Sanit. Eng. Div. Proc. ASCE **93**(SA6), 17–39 (1967)
61. Brunone, B., Ferrante, M., Romano, N., Santini, A.: Numerical simulations of one-dimensional infiltration into layered soils with the Richards equation using different estimates of the interlayer conductivity. Vadose Zone J. **2**(2), 193–200 (2003)
62. Brutsaert, W.: Probability laws for pore-size distributions. Soil Sci. **101**(2), 85–92 (1966)
63. Bués, M., Oltean, C.: Numerical simulations for saltwater intrusion by the mixed hybrid finite element method and discontinuous finite element method. Transp. Porous Media **40**(2), 171–200 (2000)
64. Burkhart, D., Hamann, B., Umlauf, G.: Adaptive and feature-preserving subdivision for high-quality tetrahedral meshes. Comput. Graph. Forum **29**(1), 117–127 (2010)
65. Burnett, R., Frind, E.: Simulation of contaminant transport in three dimensions. 1. The alternating direction Galerkin technique. Water Resour. Res. **23**(4), 683–694 (1987)
66. Caltagirone, J., Fabrie, P.: Natural convection in a porous medium at high Rayleigh numbers. Part 1 – Darcy's model. Eur. J. Mech. B/Fluids **8**, 207–227 (1989)
67. Caltagirone, J., Meyer, G., Mojtabi, A.: Structural thermoconvectives tridimensionnelles dans une couche poreuse horizontale. J. Méc. **20**, 219–232 (1981)
68. Caltagirone, J., Fabrie, P., Combarnous, M.: De la convection naturelle oscillante en milieu poreux au chaos temporel? CR Acad. Sci. Paris **305, Ser.II**, 549–553 (1987)
69. Carey, G.: Derivative calculation from finite element solutions. Comput. Methods. Appl. Mech. Eng. **35**(1), 1–14 (1982)
70. Carey, G., Barth, W., Woods, J., Kirk, B., Anderson, M., Chow, S., Bangerth, W.: Modelling error and constitutive relations in simulation of flow and transport. Int. J. Numer. Methods Fluids **46**(12), 1211–1236 (2004)

71. Carslaw, H., Jaeger, J.: Conduction of Heat in Solids. Oxford Science Publications, Oxford (1946, reprinted 2011)
72. Celia, M., Bouloutas, E., Zarba, R.: A general mass-conservative numerical solution for the unsaturated flow equation. Water Resour. Res. **26**(7), 1483–1496 (1990)
73. Chaudhry, M.: Open-Channel Flow. Prentice Hall, Englewood Cliffs (1993)
74. Chaudhry, M., Barber, M.: Open channel flow (Chapter 45). In: Johnson, R. (ed.) The Handbook of Fluid Dynamics, pp. 45:1–40. CRC/Springer, Boca Raton/Heidelberg (1998)
75. Chavent, G., Roberts, J.: A unified physical presentation of mixed, mixed-hybrid finite elements and standard finite difference approximations for the determination of velocities in waterflow problems. Adv. Water Resour. **14**(6), 329–348 (1991)
76. Cheng, R.: Modeling of hydraulic systems by finite element methods. In: Chow, V.T. (ed.) Advances in Hydroscience, vol. 11, pp. 207–283. Academic, New York (1978)
77. Cheng, P.: Heat transfer in geothermal systems. Adv. Heat Transf. **14**, 1–105 (1979)
78. Cheng, A., Quazar, D.: Analytical solutions. In: Bear, J., et al. (eds.) Seawater Intrusion in Coastal Aquifers – Concepts, Methods and Practices, pp. 163–191. Kluwer Academic, Dordrecht (1999)
79. Chilès, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty, 2nd edn. Wiley, Hoboken (2012)
80. Chilès, J.P., de Marsily, G.: Stochastic models of fracture systems and their use in flow and transport modeling. In: Bear, J., et al. (eds.) Flow and Contaminat Transport in Fractured Rock, pp. 169–236. Academic, San Diego (1993)
81. Christie, I., Mitchell, A.: Upwinding of high order Galerkin methods in conduction-convection problems. Int. J. Numer. Methods Eng. **12**(11), 1764–1771 (1978)
82. Christie, I., Griffiths, D., Mitchell, A., Zienkiewicz, O.: Finite element methods for second order differential equations with significant first derivatives. Int. J. Numer. Methods Eng. **10**(6), 1389–1396 (1976)
83. Chung, T.: Computational Fluid Dynamics. Cambridge University Press, Cambridge (2002)
84. Ciarlet, P., Lions, J.: Handbook of Numerical Analysis. Volume II. Finite Element Methods (Part 1). North-Holland, Amsterdam (1991)
85. Clausnitzer, V.: Beitrag zur Bildung und Verifikation von Parametermodellen der Mehrphasenströmung in porösen Medien (contribution to the development and verification of parametric models for multiphase flow in porous media). Master's thesis, Techn. University of Dresden, Dresden, Germany (1991)
86. Clement, T.: RT3D – a modular computer code for simulating reactive multi-species transport in 3-dimensional groundwater systems. Technical report PNNL-11720, Pacific Northwest National Laboratory, Richland (1997)
87. Clement, T., Sun, Y., Hooker, B., Petersen, J.: Modeling multi-species reactive transport in groundwater aquifers. Groundw. Monit. Remediat. J. **18**(2), 79–92 (1998)
88. Clough, R.: The finite element method in plane stress analysis. In: ASCE Structural Division. Proceedings of the 2nd Conference on Electronic Computation, Pittsburgh, pp. 345–378 (1960)
89. Cockburn, B., Gopalakrishnan, J., Wang, H.: Locally conservative fluxes for the continuous Galerkin method. SIAM J. Numer. Anal. **45**(4), 1742–1776 (2007)
90. Codina, R.: A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. Comput. Methods Appl. Mech. Eng. **110**(3–4), 325–342 (1993)
91. Codina, R.: Stability analysis of the forward Euler scheme for the convection-diffusion equation using SUPG formulation in space. Int. J. Numer. Methods Eng. **36**(9), 1445–1464 (1993)
92. Codina, R.: Comparison of some finite element methods for solving the diffusion-convection-reaction equation. Comput. Methods Appl. Mech. Eng. **156**(1–4), 185–210 (1998)
93. Codina, R.: On stabilized finite element methods for linear systems of convection-diffusion-reaction equations. Comput. Methods Appl. Mech. Eng. **188**(1–3), 61–82 (2000)

94. Coleman, B., Noll, W.: Thermodynamics of elastic materials with heat conduction and viscosity. Arch. Ration. Mech. Anal. **13**(1), 167–178 (1963)
95. Combarnous, M., Borries, S.: Hydrothermal convection in saturated porous media. In: Chow, V.T. (ed.) Advances in Hydroscience, vol. 10, pp. 231–307. Academic, New York (1975)
96. Combarnous, M., Le Fur, B.: Transfert de chaleur par convection naturelle dans une couche poreuse horizontale. CR Acad. Sci. Paris **269, Ser.B**, 1009–1012 (1969)
97. Cooper, C., Glass, R., Tyler, S.: Experimental investigation of the stability boundary for double-diffusive finger convection in a Hele-Shaw cell. Water Resour. Res. **33**(4), 517–526 (1997)
98. Cooper, C., Glass, R., Tyler, S.: Effect of buoyancy ratio on the development of double-diffusive finger convection in a Hele-Shaw cell. Water Resour. Res. **37**(9), 2323–2332 (2001)
99. Cordes, C., Kinzelbach, W.: Continuous groundwater velocity fields and path lines in linear, bilinear, and trilinear finite elements. Water Resour. Res. **28**(11), 2903–2911 (1992)
100. Corey, A.: Mechanics of immiscible fluids in porous media. Water Resources Publications, Highlands Ranch (1994)
101. Cornaton, F.: Deterministic models of groundwater age, life expectancy and transit time distributions in advective-dispersive systems. Ph.D. thesis, University of Neuchâtel, Centre of Hydrogeology, Neuchâtel (2004)
102. Cornaton, F.: Ground water – a 3-D ground water flow and transport finite element simulator. Technical report, University of Neuchâtel, Centre of Hydrogeology, Neuchâtel (2006)
103. Cornaton, F., Perrochet, P., Diersch, H.J.: A finite element formulation of the outlet gradient boundary condition for convective-diffusive transport problems. Int. J. Numer. Methods Eng. **61**(15), 2716–2732 (2004)
104. Courant, R.: Variational methods for the solution of problems of equilibrium and vibrations. Bull. Am. Math. Soc. **49**(1), 1–23 (1943)
105. Courant, R., Friedrichs, K., Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. Mathematische Annalen **100**(1), 32–74 (1928)
106. Coussy, O.: Mechanics of Porous Continua. Wiley, Chichester (1995)
107. Croucher, A., O'Sullivan, M.: The Henry problem for saltwater intrusion. Water Resour. Res. **31**(7), 1809–1814 (1995)
108. Cuthill, E., McKee, J.: Reducing the bandwidth of sparse symmetric matrices. In: Proceedings of 24th ACM National Conference, New York, pp. 157–172 (1969)
109. Dagan, G.: Stochastic modeling of flow and transport: the broad perspective. In: Dagan, G., Neuman, S. (eds.) Subsurface flow and transport: a stochastic approach, pp. 3–19. Cambridge University Press, Cambridge (1997)
110. Dahlquist, G.: A special stability problem for linear multistep methods. BIT Numer. Math. **3**(1), 27–43 (1963)
111. Danckwerts, P.: Continuous flow systems: distribution of residence times. Chem. Eng. Sci. **2**(1), 1–13 (1953)
112. Davis, S., De Wiest, R.: Hydrogeology, 2nd edn. Wiley, New York (1967)
113. Debéda, V., Caltagirone, J., Watremez, P.: Local multigrid refinement method for natural convection in fissured porous media. Numer. Heat Transf. Part B **28**(4), 455–467 (1995)
114. De Boer, R.: Theory of Porous Media. Springer, Berlin (2000)
115. De Boor, C.: A Practical Guide to Splines. Springer, New York (2001)
116. De Groot, S., Mazur, P.: Non-equilibrium Thermodynamics. Dover, Mincola (1985)
117. De Josselin de Jong, J.: Singularity distributions for the analysis of multiple-fluid flow through porous media. J. Geophys. Res. **65**(11), 3739–3758 (1960)
118. De Lemos, M.: Turbulence in porous media – modeling and applications. Elsevier, Amsterdam (2006)
119. Delleur, J.: The Handbook of Groundwater Engineering. CRC/Springer, Boca Raton (1999)
120. De Marsily, G.: Quantitative Hydrogeology – Groundwater Hydrology for Engineers. Academic, Orlando (1986)

121. Dennis, J., Moré, J.: Quasi-Newton methods, motivation and theory. SIAM Rev. **19**(1), 46–89 (1977)
122. Desai, C., Contractor, D.: Finite element analysis of flow, diffusion, and salt water intrusion in porous media. In: Bathe, K.J., et al. (eds.) Formulation and Computational Algorithms in Finite Element Analysis. MIT, Cambridge (1977)
123. Desai, C., Li, G.: A residual procedure and application for free surface flow in porous media. Adv. Water Resour. **6**(1), 27–35 (1983)
124. D'Haese, C., Putti, M., Paniconi, C., Verhoest, N.: Assessment of adaptive and heuristic time stepping for variably saturated flow. Int. J. Numer. Methods Fluids **53**(7), 1173–1193 (2007)
125. DHI-WASY: FEFLOW finite element subsurface flow and transport simulation system – User's manual/Reference manual/White papers. v. 6.1. Technical report, DHI-WASY GmbH, Berlin (2012). http://www.feflow.com
126. Diersch, H.J.: Die Berechnung stationärer zweidimensionaler und rotationssymmetrischer Potentialstrmungen mit Hilfe der Finite-Element-Methode (the computation of steady-state two-dimensional and axisymmetric potential flows by the finite element method). Wiss. Zeitschr. Techn. Univ. Dresden **24**(3/4), 801–815 (1975)
127. Diersch, H.J.: Finite-Element-Programmsystem FINEL zur Lösung von praktischen Strömungsproblemen des Wasserbaues und der Hydromechanik (finite element programming system FINEL for the solution of practical flow problems in hydraulic engineering and hydrodynamics). Wasserwirtschaft-Wassertechnik **28**(11), 385–388 (1978)
128. Diersch, H.J.: Finite-Element-Modellierung instationärer zweidimensionaler Stofftransportvorgänge im Grundwasser (finite element modeling transient two-dimensional mass transport processes in groundwater). In: Int. Konf. Simulation gekoppelter Transport-, Austausch- und Umwandlungsprozesse im Boden und Grundwasser, pp. 126–138. Tech. Univ. Dresen, Vol. 1, Dresden, Germany (1979)
129. Diersch, H.J.: Finite-element-Galerkin-Modell zur Simulation zweidimensionaler konvektiver und dispersiver Stofftransportprozesse im Boden (finite element Galerkin model for simulating convective and dispersive mass transport processes in soils). Acta Hydrophysica **26**(1), 5–44 (1981)
130. Diersch, H.J.: Primitive variables finite element solutions of free convection flows in porous media. Z. Angew. Math. Mech. **61**(7), 325–337 (1981)
131. Diersch, H.J.: On finite element upwinding and its numerical performance in simulating coupled convective transport processes. Z. Angew. Math. Mech. **63**(10), 479–488 (1983)
132. Diersch, H.J.: Modellierung und numerische Simulation geohydrodynamischer Ttransportprozesse (modeling and numerical simulation of geohydrodynamic transport processes). Ph.D. thesis, Habilitation, Academy of Sciences, Berlin, Germany (1985)
133. Diersch, H.J.: Finite element modeling of recirculating density driven saltwater intrusion processes in groundwater. Adv. Water Resour. **11**(1), 25–43 (1988)
134. Diersch, H.J.: Interactive, graphics-based finite element simulation of groundwater contamination processes. Adv. Eng. Softw. **15**(1), 1–13 (1992)
135. Diersch, H.J.: Consistent velocity approximation in the finite-element simulation of density-dependent mass and heat transport. In: FEFLOW White Papers, vol. I, Chapter 16, pp. 283–314. DHI-WASY, Berlin (2001)
136. Diersch, H.J.: The Petrov-Galerkin least square method (PGLS). In: FEFLOW White Papers, vol. I, Chapter 13, pp. 227–270. DHI-WASY, Berlin (2001)
137. Diersch, H.J., Kolditz, O.: Coupled groundwater flow and transport: 2. Thermohaline and 3D convection systems. Adv. Water Resour. **21**(5), 401–425 (1998)
138. Diersch, H.J., Kolditz, O.: Variable-density flow and transport in porous media: approaches and challenges. Adv. Water Resour. **25**(8–12), 899–944 (2002). doi:http://dx.doi.org/10.1016/S0309-1708(02)00063-5
139. Diersch, H.J., Martin, P.: Comparison of typical modelling approaches in multiple free surface, perched water table situations. In: Kovar, K., et al. (eds.) FEM MODFLOW International Conference on Finite Element Models, MODFLOW, and More: Solving Groundwater Problems, Karlovy Vary, pp. 237–240 (2004)

140. Diersch, H., Nillert, P.: Saltwater intrusion processes in groundwater: novel computer simulations, field studies and interception techniques. In: Jones, G. (ed.) International Symposium on Groundwater Monitoring and Management, Dresden, 1987. IAHS Publication, No. 173, pp. 319–329. IAHS (1990)
141. Diersch, H.J., Perrochet, P.: On the primary variable switching technique for simulating unsaturated-saturated flows. Adv. Water Resour. **23**(3), 271–301 (1999)
142. Diersch, H.J., Schirmer, A., Busch, K.F.: Analysis of flows with initially unknown discharge. J. Hydraul. Div. Proc. ASCE **103**(HY3), 213–232 (1977)
143. Diersch, H.J., Prochnow, D., Thiele, M.: Finite-element analysis of dispersion-affected saltwater upconing below a pumping well. Appl. Math. Model. **8**(5), 305–312 (1984)
144. Diersch, H.J., Clausnitzer, V., Myrnyy, V., Rosati, R., Schmidt, M., Beruda, H., Ehrnsperger, B., Virgilio, R.: Modeling unsaturated flow in absorbent swelling porous media: Part 1. Theory. Transp. Porous Media **83**(3), 437–464 (2010)
145. Diersch, H.J., Bauer, D., Heidemann, W., Rühaak, W., Schätzl, P.: Finite element modeling of borehole heat exchanger systems. Part 1. Fundamentals. Comput. Geosci. **37**(8), 1122–1135 (2011)
146. Diersch, H.J., Bauer, D., Heidemann, W., Rühaak, W., Schätzl, P.: Finite element modeling of borehole heat exchanger systems. Part 2. Numerical simulation. Comput. Geosci. **37**(8), 1136–1147 (2011)
147. Diersch, H.J., Clausnitzer, V., Myrnyy, V., Rosati, R., Schmidt, M., Beruda, H., Ehrnsperger, B., Virgilio, R.: Modeling unsaturated flow in absorbent swelling porous media: Part 2. Numerical simulation. Transp. Porous Media **86**(3), 753–776 (2011)
148. Dogrul, E., Kadir, T.: Flow computation and mass balance in Galerkin finite-element groundwater models. J. Hydraul. Eng. **132**(11), 1206–1214 (2006)
149. Donea, J., Huerta, A.: Finite Element Methods for Flow Problems. Wiley, Chichester (2003)
150. Doolen, G., Frisch, U., Hasslacher, B., Orszag, S., Wolfram, S.: Lattice Gas Methods for Partial Differential Equations. Addison-Wesley, Redwood City (1990)
151. Doyle, J.: Wave Propagation in Structures: Spectral Analysis Using Fast Discrete Fourier Transforms, 2nd edn. Springer, New York (1997)
152. Durlofsky, L.: Accuracy of mixed and control volume finite element approximations to Darcy velocity and related quantities. Water Resour. Res. **30**(4), 965–973 (1994)
153. Elder, J.: Steady free convection in a porous medium heated from below. J. Fluid Mech. **27**(1), 29–48 (1967)
154. Elder, J.: Transient convection in a porous medium. J. Fluid Mech. **27**(3), 609–623 (1967)
155. Engel, B., Navulur, K.: The role of geographical information systems in groundwater modeling (Chapter 21). In: Delleur, J. (ed.) The Handbook of Groundwater Engineering, pp. 21:1–16. CRC/Springer, Boca Raton (1999)
156. Engelman, M., Strang, G., Bathe, K.J.: The application of quasi-Newton methods in fluid mechanics. Int. J. Numer. Methods Eng. **17**(5), 707–718 (1981)
157. Eringen, A.: Mechanics of Continua, 2nd edn. Krieger, Huntington (1980)
158. Eringen, A., Ingram, J.: A continuum theory of chemically reacting media – I. Int. J. Eng. Sci. **3**(2), 197–212 (1965)
159. Eskilson, P., Claesson, J.: Simulation model for thermally interacting heat extraction boreholes. Numer. Heat Transf. **13**(2), 149–165 (1988)
160. Evans, D., Raffensperger, J.: On the stream function for variable density groundwater flow. Water Resour. Res. **28**(8), 2141–2145 (1992)
161. Fedorenko, R.: The speed of convergence of one iterative process. USSR Comput. Math. Math. Phys. **4**(3), 227–235 (1964)
162. Ferziger, J., Peric, M.: Computational Methods for Fluid Dynamics. Springer, Berlin (1996)
163. Finlayson, B.: The Method of Weighted Residuals and Variational Principles, with Applications in Fluid Mechanics, Heat and Mass Transfer. Academic, New York (1972)
164. Finlayson, B.: Numerical Methods for Problems with Moving Fronts. Ravenna Park Publishing, Seattle (1992)

165. Fletcher, C.: Computational Techniques for Fluid Dynamics, vols. 1 and 2. Springer, New York (1988)
166. Forsyth, P., Kropinski, M.: Monotonicity considerations for saturated-unsaturated subsurface flow. SIAM J. Sci. Comput. **18**(5), 1328–1354 (1997)
167. Forsyth, P., Wu, Y., Pruess, K.: Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. Adv. Water Resour. **18**(1), 25–38 (1995)
168. Forsythe, G., Wasow, W.: Finite-Difference Methods for Partial Differential Equations. Wiley, New York (1960)
169. Franca, A., Haghighi, K.: Adaptive finite element analysis of transient thermal problems. Numer. Heat Transf. Part B **26**(3), 273–292 (1994)
170. Freeze, R.: Three-dimensional transient, saturated-unsaturated flow in a ground water basin. Water Resour. Res. **7**(2), 347–366 (1971)
171. Freeze, R., Cherry, J.: Groundwater. Prentice Hall, Englewood Cliffs (1979)
172. Fries, T.P., Belytschko, T.: The extended/generalized finite element method: an overview of the method and its applications. Int. J. Numer. Methods Eng. **84**(3), 253–304 (2010)
173. Frind, E.: An isoparametric Hermitian element for the solution of field problems. Int. J. Numer. Methods Eng. **11**(6), 945–962 (1977)
174. Frind, E.: Simulation of long-term transient density-dependent transport in groundwater. Adv. Water Resour. **5**(2), 73–88 (1982)
175. Frind, E.: Solution of the advection-dispersion equation with free exit boundary. Numer. Methods Partial Differ. Equ. **4**(4), 301–313 (1988)
176. Fritsch, F., Carlson, R.: Monotone piecewise cubic interpolation. SIAM J. Numer. Anal. **17**(2), 238–246 (1980)
177. Frolkovič, P.: Consistent velocity approximation for density driven flow and transport. In: Van Keer, R., et al. (eds.) Advanced Computational Methods in Engineering, Part 2, pp. 603–611. Shaker, Maastrich (1998)
178. Frolkovič, P., De Schepper, H.: Numerical modelling of convection dominated transport with density driven flow in porous media. Adv. Water Resour. **24**(1), 63–72 (2001)
179. Fry, V., Istok, J., Guenther, R.: An analytical solution to the solute transport equation with rate-limited desorption and decay. Water Resour. Res. **29**(9), 3201–3208 (1993)
180. Galeati, G., Gambolati, G., Neuman, S.: Coupled and partially coupled Eulerian-Lagrangian model of freshwater-seawater mixing. Water Resour. Res. **28**(1), 149–165 (1992)
181. Galerkin, B.: Series solution of some problems of elastic equilibrium of rods and plates (in Russian). Vestn. Inzh. Tech. **19**, 897–908 (1915)
182. Gambolati, G., Putti, M., Paniconi, C.: Three-dimensional model of coupled density-dependent flow and miscible salt transport. In: Bear, J., et al. (eds.) Seawater Intrusion in Coastal Aquifers: Concepts, Methods and Practices, pp. 315–362. Kluwer Academic, Dordrecht (1999)
183. Garcia-Talavera, M., Laedermann, J., Decombaz, M., Daza, M., Quintana, B.: Coincidence summing corrections for the natural decay series in γ-ray spectrometry. J. Radiat. Isot. **54**, 769–776 (2001)
184. Garder, A., Jr., Peaceman, D., Pozzi, A., Jr.: Numerical simulation of multi-dimensional miscible displacement by the method of characteristics. Soc. Pet. Eng. J. **4**(1), 26–36 (1964)
185. Gardner, W.: Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. Soil Sci. **85**(4), 228–232 (1958)
186. Gartling, D., Hickox, C.: Numerical study of the applicability of the Boussinesq approximation for a fluid-saturated porous medium. Int. J. Numer. Methods Fluids **5**(11), 995–1013 (1985)
187. Gebhart, B., Jaluria, Y., Mahajan, R., Sammakia, B.: Buoyancy-Induced Flows and Transport. Hemisphere, New York (1988)
188. George, P.: Automatic Mesh Generation: Application to Finite Element Methods. Wiley, Chichester (1991)
189. George, A., Liu, J.H.: Computer Solution of Large Sparse Positive Definite Systems. Prentice Hall, Englewood Cliffs (1981)

190. Georgiadis, J., Catton, I.: Dispersion in cellular thermal convection in porous layers. Int. J. Heat Mass Transf. **31**(5), 1081–1091 (1988)
191. Ghanem, R., Spanos, P.: Stochastic Finite Elements: A Spectral Approach. Springer, New York (1991)
192. Girault, V., Raviart, P.: Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms. Springer, Berlin (1986)
193. Gockenbach, M.: Understanding and Implementing the Finite Element Method. SIAM, Philadelphia (2006)
194. Goldstein, H., Poole, C., Safko, J.: Classical Mechanics, 3rd edn. Addison Wesley, San Francisco (2002)
195. Gottardi, G., Venutelli, M.: A control-volume finite-element model for two-dimensional overland flow. Adv. Water Resour. **16**(3), 277–284 (1993)
196. Gottlieb, D., Orszag, S.: Numerical Analysis of Spectral Methods: Theory and Applications. SIAM, Philadelphia (1977)
197. Goyeau, B., Songbe, J.P., Gobin, D.: Numerical study of double-diffusive convection in a porous cavity using the Darcy-Brinkman formulation. Int. J. Heat Mass Transf. **39**(7), 1363–1378 (1996)
198. Graf, T., Therrien, R.: Variable-density groundwater flow and solute transport in porous media containing nonuniform discrete fractures. Adv. Water Resour. **28**(12), 1351–1367 (2005)
199. Granas, A., Dugundji, J.: Fixed Point Theory. Springer, New York (2003)
200. Grant, S.: Extensions of a temperature effects model for capillary pressure saturation relations. Water Resour. Res. **39**(1), 1003:1–1003:10 (2003). doi:http://dx.doi.org/10.1029/2000WR000193
201. Gray, W.: A derivation of the equations for multi-phase transport. Chem. Eng. Sci. **30**(2), 229–233 (1975)
202. Gray, W.: Derivation of vertically averaged equations describing multiphase flow in porous media. Water Resour. Res. **18**(6), 1705–1712 (1982). doi:http://dx.doi.org/10.1029/WR018i006p01705
203. Gray, W.: Thermodynamics and constitutive theory for multiphase porous-media flow considering internal geometric constraints. Adv. Water Resour. **22**(5), 521–547 (1999). doi:http://dx.doi.org/10.1016/S0309-1708(98)00021-9
204. Gray, D., Giorgini, A.: On the validity of the Boussinesq approximation for liquids and gases. Int. J. Heat Mass Transf. **19**(5), 545–551 (1976)
205. Gray, W., Hassanizadeh, S.: Paradoxes and realities in unsaturated flow theory. Water Resour. Res. **27**(8), 1847–1854 (1991). doi:http://dx.doi.org/10.1029/91WR01259
206. Gray, W., Hassanizadeh, S.: Unsaturated flow theory including interfacial phenomena. Water Resour. Res. **27**(8), 1855–1863 (1991). doi:http://dx.doi.org/10.1029/91WR01260
207. Green, T.: Scales for double-diffusive fingering in porous media. Water Resour. Res. **20**(9), 1225–1229 (1984)
208. Gresho, P., Lee, R.: Don't suppress the wiggles – they're telling you something! Comput. Fluids **9**(2), 223–253 (1981)
209. Gresho, P., Sani, R.: Incompressible flow and the finite element method. Wiley, Chichester (1998)
210. Gresho, P., Lee, R., Sani, R.: Advection-dominated flows, with emphasis on the consequences of mass lumping (Chapter 19). In: Gallagher, R., et al. (eds.) Finite Elements in Fluids, pp. 335–350. Wiley, Chichester (1978)
211. Gresho, P., Lee, R., Sani, R.: On the time-dependent solution of the incompressible Navier-Stokes equations in two and three dimensions. Technical report. Reprint UCRL-83282, Lawrence Livermore Laboratory, University of California (1979)
212. Gresho, P., Lee, R., Sani, R.: On the time-dependent solution of the incompressible Navier-Stokes equations in two and three dimensions. In: Taylor, C., Morgan, K. (eds.) Recent Advances in Numerical Methods, vol. 1, pp. 27–79. Pineridge Press, Swansea (1980)
213. Gresho, P., Lee, R., Sani, R., Maslanik, M., Eaton, B.: The consistent Galerkin FEM for computing derived boundary quantities in thermal and/or fluids problems. Int. J. Numer. Methods Fluids **7**(4), 371–394 (1987)

214. Griffiths, R.: Layered double-diffusive convection in porous media. J. Fluid Mech. **102**, 221–248 (1981)
215. Grisak, G., Pickens, J.: Solute transport through fractured media: 1. The effect of matrix diffusion. Water Resour. Res. **16**(4), 719–730 (1980)
216. Gureghian, A.: A two-dimensional finite element solution scheme for the saturated-unsaturated flow with applications to flow through ditch-drained soils. J. Hydrol. **50**, 333–353 (1981)
217. Gustaffson, I.: On first order factorization methods for the solution of problems with mixed boundary conditions and problems with discontinuous matrial coefficients. Technical report, Chalmers University of Technology and Department of Computer Sciences, University of Goeteborg, Goeteborg (1977)
218. Guymon, G., Scott, V., Herrmann, L.: A general numerical solution of the two-dimensional diffusion-convection equation by the finite element method. Water Resour. Res. **6**(6), 1611–1617 (1970)
219. Hackbusch, W.: Multi-grid Methods and Applications. Springer, Berlin (2003)
220. Häfner, F., Boy, S.: Simulation des dichteabhängigen Stofftransportes im Grundwasser und Verifizierung am Beispiel der Saltpool-Experimente (simulation of density-dependent solute transport in groundwater and verification with saltpool experiments). Grundwasser **10**(2), 93–101 (2005)
221. Häfner, F., Stüben, K.: Simulation and parameter identification of Oswald's saltpool experiments with the SAMG multigrid-solver in the transport code MODCALIF. In: Kovar, K., et al. (eds.) FEM MODFLOW International Conference on Finite Element Models, MODFLOW, and More: Solving Groundwater Problems, Karlovy Vary, pp. 23–26 (2004)
222. Hægland, H., Dahle, H., Eigestad, G., Lie, K.A., Aavatsmark, I.: Improved streamlines and time-of-flight for streamline simulation on irregular grids. Adv. Water Resour. **30**(4), 1027–1045 (2007)
223. Hansen, U., Yuen, D.: Formation of layered structures in double-diffusive convection as applied to the geosciences. In: Brandt, A., Fernando, H. (eds.) Double-Diffusive Convection. Geophysical Monograph, vol. 94, pp. 135–149. American Geophysical Union, Washington, DC (1995)
224. Hassanizadeh, S.: Modeling species transport by concentrated brine in aggregated porous media. Transp. Porous Media **3**(3), 299–318 (1988)
225. Hassanizadeh, S.: On the transient non-Fickian dispersion theory. Transp. Porous Media **23**(1), 107–124 (1996)
226. Hassanizadeh, S., Gray, W.: General conservation equations for multi-phase systems: I. Averaging procedure. Adv. Water Resour. **2**(3), 131–144 (1979). doi:http://dx.doi.org/10.1016/0309-1708(79)90025-3
227. Hassanizadeh, S., Gray, W.: General conservation equations for multi-phase systems: II. Mass, momenta, energy, and entropy equations. Adv. Water Resour. **2**(4), 191–203 (1979). doi:http://dx.doi.org/10.1016/0309-1708(79)90035-6
228. Hassanizadeh, S., Gray, W.: General conservation equations for multi-phase systems: III. Constitutive theory for porous media flow. Adv. Water Resour. **3**(1), 25–40 (1980). doi:http://dx.doi.org/10.1016/0309-1708(80)90016-0
229. Hassanizadeh, S., Gray, W.: Derivation of conditions describing transport across zones of reduced dynamics within multiphase systems. Water Resour. Res. **25**(3), 529–539 (1989)
230. Hassanizadeh, S., Gray, W.: Mechanics and thermodynamics of multiphase flow in porous media including interface boundaries. Adv. Water Resour. **13**(4), 169–186 (1990). doi:http://dx.doi.org/10.1016/0309-1708(90)90040-B
231. Hassanizadeh, S., Gray, W.: Thermodynamic basis of capillary pressure in porous media. Water Resour. Res. **29**(10), 3389–3405 (1993). doi:http://dx.doi.org/10.1029/93WR01495
232. Hassanizadeh, S., Leijnse, A.: A non-linear theory of high-concentration-gradient dispersion in porous media. Adv. Water Resour. **18**(4), 203–215 (1995). doi:http://dx.doi.org/10.1016/0309-1708(95)00012-8

233. Hassanizadeh, S., Celia, M., Dahle, H.: Dynamic effect in the capillary pressure-saturation relationship and its impacts on unsaturated flow. Vadose Zone J. **1**(1), 38–57 (2002)
234. Heidemann, W.: Zur rechnerischen Ermittlung instationärer Temperaturfelder in geschlossener und diskreter Form (on computation of transient temperature fields in closed and discrete form). Ph.D. thesis, University of Stuttgart, Stuttgart, Germany (1995)
235. Heinrich, J., Zienkiewicz, O.: Quadratic finite element schemes for two-dimensional convective-transport problems. Int. J. Numer. Methods Eng. **11**(12), 1831–1844 (1977)
236. Heinrich, J., Huyakorn, P., Zienkiewicz, O., Mitchell, A.: An 'upwind' finite element scheme for two-dimensional convective transport equation. Int. J. Numer. Methods Eng. **11**(1), 131–143 (1977)
237. Hellström, G.: Ground heat storage. Thermal analyses of duct storage systems. I. Theory. Technical report, Department of Mathematical Physics, University of Lund, Sweden (1991)
238. Helmig, R.: Multiphase Flow and Transport Processes in the Subsurface. Springer, Berlin (1997)
239. Hemker, K., Bakker, M.: Groundwater whirls in heterogeneous and anisotropic layered aquifers. In: Kovar, K., et al. (eds.) FEM MODFLOW International Conference on Finite Element Models, MODFLOW, and More: Solving Groundwater Problems, Karlovy Vary, pp. 27–30 (2004)
240. Hemker, K., Bakker, M.: Analytical solutions for whirling groundwater flow in two-dimensional heterogeneous anisotropic aquifers. Water Resour. Res. **42**(W12419), 1–12 (2006). doi:http://dx.doi.org/10.1029/2006WR004901
241. Hemker, K., van den Berg, E., Bakker, M.: Ground water whirls. Groundwater **42**(2), 234–242 (2004)
242. Henry, H.: Effects of dispersion on salt encroachment in coastal aquifers. Technical report Water-Supply Paper 1613-C, pp. 70–84, US Geological Survey (1964)
243. Henry, D., Touihri, R., Bouhlila, R., Ben Hadid, H.: Multiple flow solutions in buoyancy induced convection in a porous square box. Water Resour. Res. **48**(W10538), 1–15 (2012). doi:http://dx.doi.org/10.1029/2012WR011995
244. Herbert, A., Jackson, C., Lever, D.: Coupled groundwater flow and solute transport with fluid density strongly dependent on concentration. Water Resour. Res. **24**(10), 1781–1795 (1988)
245. Hervouet, J.M.: Hydrodynamics of Free Surface Flows. Wiley, Chichester (2007)
246. Herzberg, A.: Die Wasserversorgung einiger Nordseebäder (the water supply of parts of the North Sea coast in Germany). Z. Gasbeleucht. Wasserversorg. **44, 45**, 815–819, 842–844 (1901)
247. Hestenes, M., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. Sect. B **49**(6), 409–436 (1952)
248. Hickox, C., Gartling, D.: A numerical study of natural convection in a horizontal porous layer subjected to an end-to-end temperature difference. J. Heat Transf. **103**(4), 797–802 (1981)
249. Hills, R., Hudson, D., Porro, I., Wierenga, P.: Modeling one-dimensional infiltration into very dry soils, 1. Model development and evaluation. Water Resour. Res. **25**(6), 1259–1269 (1989)
250. Hindmarsh, A., Gresho, P., Griffiths, D.: The stability of explicit Euler time-integration for certain finite difference approximations of the multi-dimensional advection-diffusion equation. Int. J. Numer. Methods Fluids **4**(9), 853–897 (1984)
251. Hinton, E., Campbell, J.: Local and global smoothing of discontinuous finite element functions using a least squares method. Int. J. Numer. Methods Eng. **8**(3), 461–480 (1974)
252. Holst, P., Aziz, K.: Transient three-dimensional natural convection in confined porous media. Int. J. Heat Mass Transf. **15**(1), 73–90 (1972)
253. Holzbecher, E.: Modeling of saltwater upconing. In: Wang, S. (ed.) Proceedings of the 2nd International Conference Hydro-Science and Hydro-Engineering, Beijing, vol. 2, Part A, pp. 858–865 (1995)
254. Holzbecher, E.: Comment on 'constant-concentration boundary condition: lessons from the HYDROCOIN variable-density groundwater benchmark problem' by Konikow, L.F., Sanford, W.E. and Campell, P.J. Water Resour. Res. **34**(10), 2775–2778 (1998)
255. Holzbecher, E.: Modeling Density-Driven Flow in Porous Media. Springer, Berlin (1998)

256. Hood, P.: Frontal solution program for unsymmetric matrices. Int. J. Numer. Methods Eng. **10**(2), 379–399 (1976)
257. Hoopes, J., Harlemann, D.: Wastewater recharge and dispersion in porous media. J. Hydraul. Div. Proc. ASCE **93**(HY5), 51–71 (1967)
258. Horne, R.: Three-dimensional natural convection in a confined porous medium heated from below. J. Fluid Mech. **92**(4), 751–766 (1979)
259. Horne, R., Caltagirone, J.: On the evolution of thermal disturbances during natural convection in a porous medium. J. Fluid Mech. **100**(2), 385–395 (1980)
260. Horne, R., O'Sullivan, M.: Oscillatory convection in a porous medium heated from below. J. Fluid Mech. **66**(2), 339–352 (1974)
261. Horne, R., O'Sullivan, M.: Origin of oscillatory convection in a porous medium heated from below. Phys. Fluids **21**(8), 1260–1264 (1978)
262. Horton, C., Rogers, F.: Convective currents in a porous medium. J. Appl. Phys. **16**(6), 367–370 (1945)
263. Houlding, S.: 3D Geoscience Modeling. Springer, Berlin/Heidelberg (1994)
264. Hrenikoff, A.: Solution of problems in elasticity by the framework method. Trans. ASME J. Appl. Mech. **8**, 169–175 (1941)
265. Hubbert, M.: The theory of ground-water motion. J. Geol. **48**(8), 785–944 (1940)
266. Hughes, T.: A simple scheme for developing 'upwind' finite elements. Int. J. Numer. Methods Eng. **12**(9), 1359–1365 (1978)
267. Hughes, T., Franca, L.: A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. Comput. Methods Appl. Mech. Eng. **65**(1), 85–96 (1987)
268. Hughes, T., Mallet, M.: A new finite element formulation for computational fluid dynamics: III. The generalized streamline operator for multidimensional advective-diffusive systems. Comput. Methods Appl. Mech. Eng. **58**(3), 305–328 (1986)
269. Hughes, T., Mallet, M.: A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusion systems. Comput. Methods Appl. Mech. Eng. **58**(3), 329–336 (1986)
270. Hughes, J., Sanford, W.: SUTRA-MS, a version of SUTRA modified to simulate heat and multiple-solute transport. Technical report 2004-1207, p. 141, US Geological Survey, Reston (2004)
271. Hughes, J., Sanford, W., Vacher, H.: Numerical simulation of double-diffusive finger convection. Water Resour. Res. **41**(W01019), 1–16 (2005). doi:http://dx.doi.org/10.1029/2003WR002777
272. Hughes, T., Franca, L., Balestra, M.: A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuska-Brezzi condition: a stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations. Comput. Methods Appl. Mech. Eng. **59**(1), 85–99 (1986)
273. Hughes, T., Franca, L., Mallet, M.: A new finite element formulation for computational fluid dynamics: I. Symmetric forms of the compressible Euler and Navier-Sokes equations and the second law of thermodynamics. Comput. Methods Appl. Mech. Eng. **54**(2), 223–234 (1986)
274. Hughes, T., Mallet, M., Mizukami, A.: A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. Comput. Methods Appl. Mech. Eng. **54**(3), 341–355 (1986)
275. Hughes, T., Franca, L., Mallet, M.: A new finite element formulation for computational fluid dynamics: VI. Convergence analysis of the generalized SUPG formulation for linear time-dependent multidimensional advective-diffusive systems. Comput. Methods Appl. Mech. Eng. **63**(1), 97–112 (1987)
276. Hughes, T., Franca, L., Hulbert, G.: A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusion equations. Comput. Methods Appl. Mech. Eng. **73**(2), 173–189 (1989)

277. Hughes, T., Engel, G., Mazzei, L., Larson, M.: The continuous Galerkin method is locally conservative. J. Comput. Phys. **163**(2), 467–488 (2000)
278. Huyakorn, P.: Solution of steady-state, convective transport equation using an upwind finite element scheme. Appl. Math. Model. **1**(4), 187–195 (1977)
279. Huyakorn, P., Nilkuha, K.: Solution of transient transport equation using an upstream finite-element scheme. Appl. Math. Model. **3**(1), 7–17 (1979)
280. Huyakorn, P., Pinder, G.: Computational Methods in Subsurface Flow. Academic, New York (1983)
281. Huyakorn, P., Taylor, C.: Finite element models for coupled groundwater flow and convective dispersion. In: Gray, W., et al. (eds.) 1st International Conference Finite Elements in Water Resources, Princeton, pp. 1.131–1.151. Pentech, London (1976)
282. Huyakorn, P., Andersen, P., Güven, P., Molz, F.: A curvi-linear finite element model for simulating two-well tracer tests and transport in stratified aquifers. Water Resour. Res. **22**(5), 663–678 (1986)
283. Huyakorn, P., Andersen, P.F., Mercer, J., White, H., Jr.: Saltwater intrusion in aquifers: development and testing of a three-dimensional finite element model. Water Resour. Res. **23**(2), 293–312 (1987)
284. Idelsohn, S., Oñate, E.: Finite volumes and finite elements: two 'good friends'. Int. J. Numer. Methods Eng. **37**(19), 3323–3341 (1994)
285. Ingram, J., Eringen, A.: A continuum theory of chemically reacting media – II constitutive equations of reacting fluid mixtures. Int. J. Eng. Sci. **5**(4), 289–322 (1967)
286. Irmay, S.: On the hydraulic conductivity of unsaturated soil. Trans. Am. Geophys. Union **35**, 463–468 (1954)
287. Irons, B.: A frontal solution program. Int. J. Numer. Methods Eng. **2**(1), 5–32 (1970)
288. Johannsen, K.: On the validity of the Boussinesq approximation for the Elder problem. Comput. Geosci. **7**(3), 169–182 (2003)
289. Johannsen, K., Kinzelbach, W., Oswald, S., Wittum, G.: The saltpool benchmark problem – numerical simulation of saltwater upconing in a porous medium. Adv. Water Resour. **25**(3), 335–348 (2002)
290. Johns, R., Rivera, A.: Comment on 'dispersive transport dynamics in a strongly coupled groundwater-brine flow system' by Oldenburg, C.M. and Pruess, K. Water Resour. Res. **32**(11), 3405–3410 (1996)
291. Johnson, C.: Numerical Solution of Partial Differential Equations by the Finite Element Method. Cambridge University Press, Cambridge (1987)
292. Johnson, C., Szepessy, A., Hansbo, P.: On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservations laws. Math. Comput. **54**(189), 107–129 (1990)
293. Josnin, J.Y., Jourde, H., Fénart, P., Bidaux, P.: A three-dimensional model to simulate joint networks in layered rocks. Can. J. Earth Sci. **39**(10), 1443–1455 (2002)
294. Jourde, H., Cornaton, F., Pistre, S., Bidaux, P.: Flow behavior in a dual fracture network. J. Hydrol. **266**(1–2), 99–119 (2002)
295. Ju, S.H., Kung, K.J.: Mass types, element orders and solution schemes for the Richards equation. Comput. Geosci. **23**(2), 175–187 (1997)
296. Kakaç, S., Kilkiş, B., Kulacki, F., Arinç, F. (eds.): Convective Heat and Mass Transfer in Porous Media. NATO ASI Series. Kluwer Academic, Dordrecht (1991)
297. Kaluarachchi, J., Parker, J.: An efficient finite element method for modeling multiphase flow. Water Resour. Res. **25**(1), 43–54 (1989)
298. Kämpf, M., Holfelder, T., Montenegro, H.: Identification and parametrization of flow processes in artificial capillary barriers. Water Resour. Res. **39**(10,1276), 1–9 (2003). doi:http://dx.doi.org/10.1029/2002WR001860
299. Kanney, J., Miller, C., Kelley, C.: Convergence of iterative split-operator approaches for approximating nonlinear reactive transport problems. Adv. Water Resour. **26**(3), 247–261 (2003)

300. Kantorovich, L.: A direct method of solving problems on the minimum of a double integral. Bull. Acad. Sci. USSR **5**, 647–652 (1933)
301. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. **20**(1), 359–392 (1998)
302. Karypis, G., Kumar, V.: METIS – a software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. v. 4.0. Technical report, Department of Computer Science, University of Minnesota, Minnesota (1998). http://www.cs.umn.edu/∼metis
303. Kastanek, F., Nielsen, D.: Description of soil water characteristics using cubic spline interpolation. Soil Sci. Soc. Am. J. **65**(2), 279–283 (2001)
304. Katto, Y., Masuoka, I.: Criterion for the onset of convective flow in a fluid in a porous medium. Int. J. Heat Mass Transf. **10**(3), 297–309 (1967)
305. Kaviany, M.: Principles of Heat Transfer in Porous Media, 2nd edn. Springer, New York (1995)
306. Kazemi, G., Lehr, J., Perrochet, P.: Groundwater Age. Wiley-Interscience, Hoboken (2006)
307. Kelly, D., Nakazawa, S., Zienkiewicz, O., Heinrich, J.: A note on upwinding and anisotropic balancing dissipation in finite element approximations to convective diffusion problems. Int. J. Numer. Methods Eng. **15**(11), 1705–1711 (1980)
308. Kimura, S., Schubert, G., Straus, J.: Route to chaos in porous-medium thermal convection. J. Fluid Mech. **166**, 23–32 (1986)
309. Kimura, S., Schubert, G., Straus, J.: Instabilities of steady, periodic and quasi-periodic modes of convection in porous media. ASME J. Heat Transf. **109**(2), 350–355 (1987)
310. Kinzelbach, W.: Groundwater Modelling: An Introduction with Sample Programs in BASIC. Elsevier, Amsterdam (1986)
311. Kirkland, M., Hills, R., Wierenga, P.: Algorithms for solving Richards' equation for variably saturated soils. Water Resour. Res. **28**(8), 2049–2058 (1992)
312. Knabner, P., Frolkovič, P.: Consistent velocity approximation for finite volume or element discretizations of density driven flow in porous media. In: Aldama, A., et al. (eds.) Computational Methods in Water Resources XI – Computational Methods in Subsurface Flow and Transport Problems, vol. 1, pp. 93–100. Computational Mechanics Publications, Southampton (1996)
313. Knupp, P.: A moving mesh algorithm for 3-D regional groundwater flow with water table and seepage face. Adv. Water Resour. **19**(2), 83–95 (1996)
314. Knupp, P., Steinberg, S.: Fundamentals of Grid Generation. CRC, Boca Raton (1994)
315. Knuth, D.: The Art of Computer Programming, vol. 3: Sorting and Searching. Addison-Wesley, Reading (1997)
316. Kolditz, O.: Strömung, Stoff- und Wärmetransport im Kluftgestein (flow, mass and heat transport in fractured rock). Gebr. Borntraeger, Berlin/Stuttgart (1997)
317. Kolditz, O.: Computational Methods in Environmental Fluid Mechanics. Springer, Berlin (2002)
318. Kolditz, O., Ratke, R., Diersch, H.J., Zielke, W.: Coupled groundwater flow and transport: 1. Verification of variable-density flow and transport models. Adv. Water Resour. **21**(1), 27–46 (1998)
319. König, C.: Operator split for three dimensional mass transport equation. In: Peters, A., et al. (eds.) Proceedings of the 10th International Conference on Computational Methods in Water Resources, Heidelberg, vol. 1, pp. 309–316. Kluwer Academic, Dordrecht (1994)
320. Konikow, L., Sanford, W., Campbell, P.: Constant-concentration boundary condition: lessons from the HYDROCOIN variable-density groundwater benchmark problem. Water Resour. Res. **33**(10), 2253–2261 (1997)
321. Kool, J., Parker, J.: Development and evaluation of closed-form expressions for hysteretic soil hydraulic properties. Water Resour. Res. **23**(1), 105–114 (1987)
322. Krieger, R., Hatchett, J., Poole, J.: Preliminary survey of the saline-water resources of the United States. Technical report Water-Supply Paper 1374, p. 172, US Geological Survey (1957)

323. Kvernvold, O., Tyvand, P.: Nonlinear thermal convection in anisotropic porous media. J. Fluid Mech. **90**(4), 609–624 (1979)
324. Kvernvold, O., Tyvand, P.: Dispersion effects on thermal convection in porous media. J. Fluid Mech. **99**(4), 673–686 (1980)
325. Kvernvold, O., Tyvand, P.: Dispersion effects on thermal convection in a Hele-Shaw cell. Int. J. Heat Mass Transf. **24**(5), 887–890 (1981)
326. Labbé, P., Garon, A.: A robust implementation of Zienkiewicz and Zhu's local patch recovery method. Commun. Numer. Methods Eng. **11**(5), 427–434 (1995)
327. LaBolle, E., 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. **35**(2), 605–607 (1999)
328. Lacombe, S., Sudicky, E., Frape, S., Unger, A.: Influence of leaky boreholes on cross-formational groundwater flow and contaminant transport. Water Resour. Res. **31**(8), 1871–1882 (1995)
329. Lam, L., Fredlund, D.: Saturated-unsaturated transient finite element seepage model for geotechnical engineering. Adv. Water Resour. **7**(3), 132–136 (1984)
330. Lamarche, L., Kajl, S., Beauchamp, B.: A review of methods to evaluate borehole thermal resistances in geothermal heat-pump systems. Geothermics **39**(2), 187–200 (2010)
331. Lambert, J.: Computational Methods in Ordinary Differential Equations. Wiley, London (1973)
332. Lanczos, C.: Solutions of systems of linear equations by minimized iterations. J. Res. Natl. Bur. Stand. Sect. B **49**(1), 33–53 (1952)
333. Lapwood, E.: Convection of a fluid in a porous medium. Math. Proc. Camb. Phil. Soc. **44**(4), 508–521 (1948)
334. Lee, U.: Spectral Element Method in Structural Dynamics. Wiley, Singapore (2009)
335. Lehmann, F., Ackerer, P.: Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media. Transp. Porous Media **31**(3), 275–292 (1998)
336. Leijnse, A.: Three-dimensional modeling of coupled flow and transport in porous media. Ph.D. thesis, University of Notre Dame, Indiana (1992)
337. Leijnse, A.: Comparison of solution methods for coupled flow and transport in porous media. In: Peters, A., et al. (eds.) Proceedings of the 10th International Conference on Computational Methods in Water Resources, Heidelberg, vol. 1, pp. 489–496. Kluwer Academic, Dordrecht (1994)
338. Leismann, H., Frind, E.: A symmetric-matrix time integration scheme for the efficient solution of advection-dispersion problems. Water Resour. Res. **25**(6), 1133–1139 (1989)
339. Lenhard, R., Parker, J.: A model for hysteretic constitutive relations governing multiphase flow. 2. Permeability-saturation relations. Water Resour. Res. **23**(12), 2197–2206 (1987)
340. Lenhard, R., Parker, J., Kaluarachchi, J.: Comparing simulated and experimental hysteretic two-phase transient fluid flow phenomena. Water Resour. Res. **27**(8), 2113–2124 (1991)
341. Leone, J., Gresho, P., Chan, S., Lee, R.: A note on the accuracy of Gauss-Legendre quadrature in the finite element method. Int. J. Numer. Methods Eng. **14**(5), 769–773 (1979)
342. Letniowski, F.: An overview of preconditioned iterative methods for sparse matrix equations. Technical report, CS-89-26, Faculty of Mathematics, University of Waterloo, Waterloo (1989)
343. Lever, D., Jackson, C.: On the equations for the flow of concentrated salt solution through a porous medium. Technical report, AERE-R 11765, Harwell Laboratory, Oxfordshire (1985)
344. Lewis, R., Schrefler, B.: The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media, 2nd edn. Wiley, Chichester (1998)
345. Lewis, R., Morgan, K., Thomas, H., Seetharamu, K.: The Finite Element Method in Heat Transfer Analysis. Wiley, Chichester (1996)
346. Li, W., Chen, Z., Ewing, R., Huan, G., Li, B.: Comparison of the GMRES and ORTHOMIN for the black oil model in porous media. Int. J. Numer. Methods Fluids **48**(5), 501–519 (2005)
347. Lian, Y.Y., Hsu, K.H., Shao, Y.L., Lee, Y.M., Jeng, Y.W., Wu, J.S.: Parallel adaptive mesh-refining scheme on a three-dimensional unstructured tetrahedral mesh and its applications. Comput. Phys. Commun. **175**(11–12), 721–737 (2006)

348. Lichtner, P.: Continuum formulation of multicomponent-multiphase reactive transport. In: Lichtner, P., et al. (eds.) Reactive Transport in Porous Media. Reviews in Mineralogy, vol. 34, pp. 1–81. Mineralogical Society of America, Washington, DC (1996)
349. Lichtner, P., Kelkar, S., Robinson, B.: Critique of Burnett-Frind dispersion tensor for axisymmetric porous media. Technical report LA-UR-08-04495, Los Alamos National Laboratory (2008)
350. Liggett, J., Liu, P.F.: The Boundary Integral Equation Method for Porous Media Flow. Unwin Hyman, Boston (1982)
351. Lindstrom, F.: Pulsed dispersion of trace chemical concentrations in a saturated sorbing porous medium. Water Resour. Res. **12**(2), 229–238 (1976)
352. Liu, G.: Mesh Free Methods: Moving Beyond the Finite Element Method. CRC, Boca Raton (2003)
353. Löhner, R.: Applied CFD Techniques. Wiley, Chichester (2001)
354. Löhner, R., Morgan, K.: An unstructured multigrid method for elliptic problems. Int. J. Numer. Methods Eng. **24**(1), 101–115 (1987)
355. Lynch, D.: Mass conservation in finite element groundwater models. Adv. Water Resour. **7**(2), 67–75 (1984)
356. Maidment, D.: Handbook of Hydrology. McGraw-Hill, New York (1993)
357. Matheron, G.: The intrinsic random functions and their applications. Adv. Appl. Probab. **5**, 439–468 (1973)
358. Matthews, C., Cook, F., Knight, J., Braddock, R.: Handling the water content discontinuity at the interface between layered soils with a numerical scheme. In: SuperSoil 2004 – 3rd Australian New Sealand Soils Conference, University of Sydney, Australia, pp. 1–9, 5–9 Dec 2004
359. Mazzia, A., Putti, M.: High order Godunov mixed methods on tetrahedral meshes for density driven flow simulations in porous media. J. Comput. Phys. **208**(1), 154–174 (2005)
360. Mazzia, A., Bergamaschi, L., Putti, M.: On the reliability of numerical solutions of brine transport in groundwater: analysis of infiltration from a salt lake. Transp. Porous Media **43**(1), 65–86 (2001)
361. McBride, D., Cross, M., Croft, N., Bennett, C., Gebhardt, J.: Computational modelling of variably saturated flow in porous media with complex three-dimensional geometries. Int. J. Numer. Methods Fluids **50**(9), 1085–1117 (2006)
362. McCord, J.: Application of second-type boundaries in unsaturated flow modeling. Water Resour. Res. **27**(12), 3257–3260 (1991)
363. McDonald, M., Harbaugh, A.: A modular three-dimensional finite-difference ground-water flow model. Technical report, Open-File Report 83–875, U.S. Geological Survey (1988)
364. McKibbin, R., O'Sullivan, M.: Onset of convection in a layered porous medium heated from below. J. Fluid Mech. **96**(2), 375–393 (1980)
365. McKibbin, R., O'Sullivan, M.: Heat transfer in layered porous medium heated from below. J. Fluid Mech. **111**, 141–173 (1981)
366. McKibbin, R., Tyvand, P.: Anisotropic modelling of thermal convection in multilayered porous media. J. Fluid Mech. **118**, 315–339 (1982)
367. McKibbin, R., Tyvand, P.: Thermal convection in a porous medium composed of alternating thick and thin layers. Int. J. Heat Mass Transf. **26**(5), 761–780 (1983)
368. McKibbin, R., Tyvand, P.: Thermal convection in a porous medium with horizontal cracks. Int. J. Heat Mass Transf. **27**(7), 1007–1023 (1984)
369. McLaren, R.: GRIDBUILDER: a generator for 2D triangular finite element grids and grid properties – User's guide. Technical report, Waterloo Institute for Groundwater Research, University of Waterloo, Waterloo (1995). http://www.science.uwaterloo.ca/~mclaren/
370. Meesters, A., Hemker, C., van den Berg, E.: An approximate analytical solution for well flow in anisotropic layered aquifer systems. J. Hydrol. **296**(1–4), 241–253 (2004)
371. Mercer, J., Pinder, G.: Finite element analysis of hydrothermal systems. In: Oden, J., et al. (eds.) Finite Element Methods in Flow Problems. Proceedings of 1st Symposium, Swansea, pp. 401–414. University of Alabama Press, Huntsville (1974)

372. Merchant, M., Weatherill, N.: Adaptivity techniques for compressible inviscid flows. Comput. Methods Appl. Mech. Eng. **106**(1–2), 83–106 (1993)
373. Milly, P.: A mass-conservative procedure for time-stepping in models of unsaturated flow. Adv. Water Resour. **8**(1), 32–36 (1985)
374. Minkowycz, W., Sparrow, E., Murthy, J.: Handbook of Numerical Heat Transfer, 2nd edn. Wiley, Hoboken (2006)
375. Mirnyy, V., Clausnitzer, V., Diersch, H.J., Rosati, R., Schmidt, M., Beruda, H.: Wicking in absorbent swelling porous materials (Chapter 7). In: Masoodi, R., Pillai, K. (eds.) Wicking in Porous Materials, pp. 161–200. CRC/Taylor and Francis, Boca Raton (2013)
376. Mitchell, A., Griffiths, R.: The Finite Difference Method in Partial Differential Equations. Wiley, Chichester (1980)
377. Mitchell, A., Wait, R.: The Finite Element Method in Partial Differential Equations. Wiley, Chichester (1977)
378. 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**(11), 3001–3012 (1994)
379. Mualem, Y.: A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour. Res. **12**(3), 513–521 (1976). doi:http://dx.doi.org/10.1029/WR012i003p00513
380. Murray, B., Chen, C.: Double-diffusive convection in a porous medium. J. Fluid Mech. **201**, 147–166 (1989)
381. Muskat, M.: The Flow of Homogeneous Fluids Through Porous Media. McGraw-Hill, New York (1937). Reprinted by J.W. Edwards, Ann Arbor, 1946
382. Narasimhan, T., Whitherspoon, P.: Numerical model for saturated-unsaturated flow in deformable porous media, 3, applications. Water Resour. Res. **14**(6), 1017–1034 (1978)
383. Neuman, S.: Saturated-unsaturated seepage by finite elements. J. Hydraul. Div. Proc. ASCE **99**(HY12), 2233–2250 (1973)
384. Neuman, S., Witherspoon, P.: Analysis of nonsteady flow with a free surface using the finite element method. Water Resour. Res. **7**(3), 611–623 (1971)
385. Nguyen, H., Reynen, J.: A space-time least-square finite element scheme for advection-diffusion equations. Comput. Methods Appl. Mech. Eng. **42**(3), 331–342 (1984)
386. Nguyen, V., Gray, W., Pinder, G., Botha, J., Crerar, D.: A theoretical investigation on the transport of chemicals in reactive porous media. Water Resour. Res. **18**(4), 1149–1156 (1982)
387. Nield, D.: Onset of thermohaline convection in a porous medium. Water Resour. Res. **4**(3), 553–560 (1968)
388. Nield, D.: The stability of convective flows in porous media. In: Kakaç, S., et al. (eds.) Convective Heat and Mass Transfer in Porous Media. NATO ASI Series, pp. 79–122. Kluwer Academic, Dordrecht (1991)
389. Nield, D., Bejan, A.: Convection in Porous Media, 3rd edn. Springer, New York (2006)
390. Nield, D., Simmons, C., Kuznetsov, A., Ward, J.: On the evolution of salt lakes: episodic convection beneath an evaporating salt lake. Water Resour. Res. **44**(W02439), 1–13 (2008). doi:http://dx.doi.org/10.1029/2007WR006161
391. Nillert, P.: Beitrag zur Simulation von Brunnen als innere Randbedingungen in horizontalebenen diskreten Grundwasserströmungsmodellen (simulation of wells as inner boundary conditions for horizontal 2D discrete groundwater flow models). Ph.D. thesis, Technical University Dresden, Dresden, Germany (1976)
392. Nordbotten, J., Celia, M., Dahle, H., Hassanizadeh, S.: Interpretation of macroscale variables in Darcy's law. Water Resour. Res. **43**(W08430), 1–9 (2007). doi:http://dx.doi.org/10.1029/2006WR005018
393. Oberbeck, A.: Über die Wärmeleitung der Flüssigkeiten bei Berücksichtigung der Strömung infolge von Temperaturdifferenzen (on the thermal conduction of liquids with regard to flows due to temperature differences). Ann. Phys. Chem. **7**, 271–292 (1879)
394. Ochoa-Tapia, J., Whitaker, S.: Momentum transfer at the boundary between a porous medium and a homogeneous fluid – I. Theoretical development. Int. J. Heat Mass Transf. **38**(14), 2635–2646 (1995)

395. OECD: The international INTRAVAL project, Phase 1, Summary report. Technical report, OECD, Paris (1994)
396. Ogata, A., Banks, R.: A solution of the differential equation of longitudinal dispersion in porous media. Technical report professional paper 411-A, A1-A9, US Geological Survey (1961)
397. Oldenburg, C., Pruess, K.: On numerical modeling of capillary barriers. Water Resour. Res. **29**(4), 1045–1056 (1993)
398. Oldenburg, C., Pruess, K.: Dispersive transport dynamics in a strongly coupled groundwater-brine flow system. Water Resour. Res. **31**(2), 289–302 (1995)
399. Oldenburg, C., Pruess, K.: Layered thermohaline convection in hypersaline geothermal systems. Transp. Porous Media **33**(1/2), 29–63 (1998)
400. Oldenburg, C., Pruess, K., Travis, B.: Reply to: comment on 'dispersive transport dynamics in a strongly coupled groundwater-brine flow system' by Johns, R.T. and Rivera, A. Water Resour. Res. **32**(11), 3411–3412 (1996)
401. Oltean, C., Bués, M.: Coupled groundwater flow and transport in porous media. A conservative or non-conservative form? Transp. Porous Media **44**(2), 219–246 (2001)
402. Oñate, E., Bugeda, G.: Mesh optimality criteria for adaptive finite element computations. In: Whiteman, J. (ed.) The Mathematics of Finite Elements and Applications – Highlights 1993, pp. 121–135. Wiley, Chichester (1994)
403. Oshima, M., Hughes, T., Jansen, K.: Consistent finite element calculation of boundary and internal fluxes. Int. J. Comput. Fluid Dyn. **9**(3–4), 227–235 (1998)
404. Oswald, S.: Dichteströmungen in porösen Medien: Dreidimensionale Experimente und Modellierungen (density driven flows in porous media: three-dimensional experiments and modeling). Ph.D. thesis, ETH Zurich, Switzerland (1998)
405. Oswald, S., Kinzelbach, W.: A three-dimensional physical model for verification of variable-density flow codes. In: Stauffer, F., et al. (eds.) MODELCARE99 – Calibration and Reliability in Groundwater Modelling: Coping with Uncertainty, Zurich, 1999. IAHS Publication, No. 265, pp. 399–404. IAHS (2000)
406. Oswald, S., Kinzelbach, W.: Three-dimensional physical benchmark experiments to test variable-density flow models. J. Hydrol. **290**(1–2), 22–42 (2004)
407. Panday, S., Forsyth, P., Falta, R., Wu, Y.S., Huyakorn, P.: Considerations of robust compositional simulations of subsurface nonaquaeous phase liquid contamination and remediation. Water Resour. Res. **31**(5), 1273–1289 (1995)
408. Paniconi, C., Putti, M.: A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems. Water Resour. Res. **30**(12), 3357–3374 (1994)
409. Panton, R.: Incompressible Flow. Wiley, New York (1996)
410. Park, C.H., Aral, M.: Sensitivity of the solution of the Elder problem to density, velocity and numerical perturbations. J. Contam. Hydrol. **92**(1–2), 33–49 (2007)
411. Parker, J., Lenhard, R.: A model for hysteretic constitutive relations governing multiphase flow. 1. Saturation-pressure relations. Water Resour. Res. **23**(12), 2187–2196 (1987)
412. Pasdunkorale, J., Turner, I.: A second order finite volume technique for simulating transport in anisotropic media. Int. J. Numer. Methods Heat Fluid Flow **13**(1), 31–56 (2003)
413. Pasquetti, R., Rapetti, F.: Spectral element methods on triangles and quadrilaterals: comparisons and applications. J. Comput. Phys. **198**(1), 349–362 (2004)
414. Perrochet, P.: Finite hyperelements: a 4D geometrical framework using covariant bases and metric tensors. Commun. Numer. Methods Eng. **11**(6), 525–534 (1995)
415. Perrochet, P., Bérod, D.: Stability of the standard Crank-Nicolson-Galerkin scheme applied to the diffusion-convection equation: some new insights. Water Resour. Res. **29**(9), 3291–3297 (1993)
416. Peters, A., Durner, W., Wessolek, G.: Consistent parameter constraints for soil hydraulic functions. Adv. Water Resour. **34**(10), 1352–1365 (2011)
417. Philip, J.: Theory of infiltration. In: Chow, V.T. (ed.) Advances in Hydroscience, vol. 5, pp. 215–296. Academic, New York (1969)

418. Piessanetzky, S.: Sparse Matrix Technology. Academic, New York (1984)
419. Pinder, G.: Groundwater Modeling Using Geographical Information Systems. Wiley, New York (2002)
420. Pinder, G., Cooper, H.: A numerical technique for calculating the transient position of the saltwater front. Water Resour. Res. **6**(3), 875–882 (1970)
421. Pinder, G., Gray, W.: Finite Element Simulation in Surface and Subsurface Hydrology. Academic, New York (1977)
422. Pinder, G., Gray, W.: Essentials of Multiphase Flow and Transport in Porous Media. Wiley, Hoboken (2008)
423. Pironneau, O.: Finite Element Methods for Fluids. Wiley, New York (1989)
424. Pokrajac, D., Lazic, R.: An efficient algorithm for high accuracy particle tracking in finite elements. Adv. Water Resour. **25**(4), 353–369 (2002)
425. Pollock, D.: Semianalytical computation of path lines for finite-difference models. Groundwater **26**(6), 743–750 (1988)
426. Polubarinova-Kochina, P.: Theory of Groundwater Movement. Princeton University Press, Princeton (1962)
427. Prakash, A.: Finite element solutions of the non-self-adjoint convective-dispersion equation. Int. J. Numer. Methods Eng. **11**(2), 269–287 (1977)
428. Prasad, V., Kladias, N.: Non-Darcy natural convection in saturated porous media. In: Kakaç, S., et al. (eds.) Convective Heat and Mass Transfer in Porous Media. NATO ASI Series, pp. 173–224. Kluwer Academic, Dordrecht (1991)
429. Prasad, A., Simmons, C.: Unstable density-driven flow in heterogeneous media: a stochastic study of the Elder [1967b] 'short heater' problem. Water Resour. Res. **39**(1), 107 (2003). doi:http://dx.doi.org/10.1029/2002WR001290
430. Press, W., Flannery, B., Teukolsky, S., Vetterling, W.: Numerical Recipes in C. Cambridge University Press, Cambridge (1988)
431. Pringle, S., Glass, R., Cooper, C.: Double-diffusive finger convection in a Hele-Shaw cell: an experiment exploring the evolution of concentration fields, length scales and mass transfer. Transp. Porous Media **47**(2), 195–214 (2002)
432. Putti, M., Paniconi, C.: Picard and Newton linearization for the coupled model of saltwater intrusion in aquifer. Adv. Water Resour. **18**(3), 159–170 (1995)
433. Rannacher, R., Bangerth, W.: Adaptive Finite Element Methods for Differential Equations. Birkhäuser, Basel (2003)
434. Raper, J.: Three Dimensional Applications in Geographic Information Systems. Taylor and Francis, London (1989)
435. Rathfelder, K., Abriola, L.: Mass conservative numerical solutions of the head-based Richards equation. Water Resour. Res. **30**(9), 2579–2586 (1994)
436. Raviart, P., Thomas, J.: A mixed finite element method for the second order elliptic problems. In: Galligani, I., Magenes, E. (eds.) Mathematical Aspects of Finite Element Methods. Lecture Notes in Mathematics, vol. 606, pp. 292–315. Springer, Berlin (1977)
437. Reddy, S., Gartling, D.: The Finite Element Method in Heat Transfer and Fluid Dynamics, 2nd edn. CRC, Boca Raton (2001)
438. Reilly, T., Goodman, A.: Quantitative analysis of saltwater-freshwater relationships in groundwater systems – a historical perspective. J. Hydrol. **80**(1–2), 125–160 (1985)
439. Reilly, T., Goodman, A.: Analysis of saltwater upconing beneath a pumping well. J. Hydrol. **89**(3–4), 169–204 (1987)
440. Richards, L.: Capillary conduction of liquids through porous media. Physics **1**, 318–333 (1931)
441. Richtmeyer, R., Morton, K.: Difference Methods for Initial Value Problems, 2nd edn. Wiley-Interscience, New York (1963)
442. Rifai, S., Newell, C., Miller, C., Taffinder, S., Rounsaville, M.: Simulation of natural attenuation with multiple electron acceptors. Bioremediation **3**(1), 53–58 (1995)
443. Rijtema, P.: An analysis of actual evapotranspiration. Technical report 659, p. 170, Center for Agricultural Publishing and Documentation, Wageningen (1965)

444. Riley, D., Winters, K.: Modal exchange mechanisms in Lapwood convection. J. Fluid Mech. **204**, 325–358 (1989)
445. Ritz, W.: Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik. J. Reine Angew. Math. **135**, 1–61 (1908)
446. Roache, P.: On artificial viscosity. J. Comput. Phys. **10**(2), 169–184 (1972)
447. Rohsenow, W., Hartnett, J., Cho, Y.: Handbook of Heat Transfer, 3rd edn. McGraw-Hill, New York (1998)
448. Ross, B.: The diversion capacity of capillary barriers. Water Resour. Res. **26**(10), 2625–2629 (1990)
449. Ross, P., Bistow, K.: Simulating water movement in layered and gradational soils using the Kirchhoff transform. Soil Sci. Soc. Am. J. **54**(6), 1519–1524 (1990)
450. Rubin, H.: Onset of thermohaline convection in a cavernous aquifer. Water Resour. Res. **12**(2), 141–147 (1976)
451. Rubin, H., Roth, C.: On the growth of instabilities in groundwater due to temperature and salinity gradients. Adv. Water Resour. **2**, 69–76 (1979)
452. Rubin, H., Roth, C.: Thermohaline convection in flowing groundwater. Adv. Water Resour. **6**(3), 146–156 (1983)
453. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. SIAM - Society for Industrial and Applied Mathematics, Philadelphia (2003)
454. Saad, Y., Schultz, M.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**(3), 856–869 (1986)
455. Sadek, E.: A scheme for the automatic generation of triangular finite elements. Int. J. Numer. Methods Eng. **15**(12), 1813–1822 (1980)
456. Salvadori, M., Baron, M.: Numerical Methods in Engineering. Prentice Hall, Englewood Cliffs (1961)
457. Sarler, B., Gobin, D., Goyeau, B., Perko, J., Power, H.: Natural convection in porous media – dual reciprocity boundary element method solution of the Darcy model. Int. J. Numer. Methods Fluids **33**(2), 279–312 (2000)
458. Schätzl, P., Clausnitzer, V., Diersch, H.J.: Groundwater modeling for mining and underground construction – challenges and solutions. In: MODFLOW and More: Ground Water and Public Policy, Golden, pp. 58–61. International Ground Water Modeling Center (IGWMC), Colorado School of Mines, US (2008)
459. Scheidegger, A.: The Physics of Flow Through Porous Media. MacMillan, New York (1957)
460. Scheidegger, A.: General theory of dispersion in porous media. J. Geophys. Res. **66**(10), 3273–3278 (1961)
461. Schenk, O., Gärtner, K.: Solving unsymmetric sparse systems of linear equations with PARDISO. J. Future Gener. Comput. Syst. **20**(3), 475–487 (2004)
462. Schiesser, W.: The Numerical Method of Lines: Integration of Partial Differential Equations. Academic, San Diego (1991)
463. Schneider, K.: Investigation on the influence of free thermal convection on heat transfer through granular material. In: Proceedings of the 11th International Congress of Refrigeration, Paper 11-4, Munich, pp. 247–253. Pergamon, Oxford (1963)
464. Schotting, R., Moser, H., Hassanizadeh, S.: High-concentration-gradient dispersion in porous media: experiments, analysis and approximations. Adv. Water Resour. **22**(7), 665–680 (1999). doi:http://dx.doi.org/10.1016/S0309-1708(98)00052-9
465. Schubert, G., Straus, J.: Three-dimensional and multicellular steady and unsteady convection in fluid-saturated porous media at high Rayleigh numbers. J. Fluid Mech. **94**(1), 25–38 (1979)
466. Schubert, G., Straus, J.: Transitions in time-dependent thermal convection in fluid-saturated porous media. J. Fluid Mech. **121**, 301–313 (1982)
467. Schulz, R.: Analytical model calculations for heat exchange in a confined aquifer. J. Geophys. **61**, 12–20 (1987)
468. Schwarz, H.: Methode der finiten Elemente (Method of Finite Elements). B.G. Teubner, Stuttgart (1991)

469. Scott, P., Farquhar, G., Kouwen, N.: Hysteresis effects on net infiltration. In: Advances in Infiltration. Publication 11–83, pp. 163–170. American Society of Agricultural Engineers, St. Joseph (1983)
470. Segol, G.: Classic Groundwater Simulations: Proving and Improving Numerical Models. Prentice Hall, Englewood Cliffs (1994)
471. Segol, G., Pinder, G.: Transient simulation of saltwater intrusion in southeastern Florida. Water Resour. Res. **12**(1), 65–70 (1976)
472. Segol, G., Pinder, G., Gray, W.: A Galerkin-finite element technique for calculating the transient position of the saltwater front. Water Resour. Res. **11**(2), 343–347 (1975)
473. Selker, J., Keller, C., McCord, J.: Vadose Zone Processes. Lewis, Boca Raton (1999)
474. Shamsai, A., Narasimhan, T.: A numerical investigation of free surface-seepage face relationship under steady state flow conditions. Water Resour. Res. **27**(3), 409–421 (1991)
475. Shewchuk, J.: TRIANGLE: a two-dimensional quality mesh generator and Delaunay triangulator. Technical report, University of California, Computer Science Division, Berkeley (2005). https://www.cs.cmu.edu/~quake/triangle.html
476. Siegel, P., Mosé, R., Ackerer, P., Jaffre, J.: Solution of the advection-diffusion equation using a combination of discontinuous and mixed finite elements. Int. J. Numer. Methods Fluids **24**(6), 595–613 (1997)
477. Signorelli, S., Bassetti, S., Pahud, D., Kohl, T.: Numerical evaluation of thermal response tests. Geothermics **36**(2), 141–166 (2007)
478. Simmons, C.: Variable density groundwater flow: from current challenges to future possibilities. Hydrogeol. J. **13**(1), 116–119 (2005)
479. Simmons, C., Narayan, K., Wooding, R.: On a test case for density-dependent groundwater flow and solute transport models: the salt lake problem. Water Resour. Res. **35**(12), 3607–3620 (1999)
480. Simmons, C., Fenstemaker, T., Sharp, J., Jr.: Variable-density flow and solute transport in heterogeneous porous media: approaches, resolutions and future challenges. J. Contam. Hydrol. **52**(1–4), 245–275 (2001)
481. Simmons, C., Pierini, M., Hutson, J.: Laboratory investigation of variable-density flow and solute transport in unsaturatedsaturated porous media. Transp. Porous Media **47**(2), 215–244 (2002)
482. Simpson, M., Clement, T.: Improving the worthiness of the Henry problem as a benchmark for density-dependent groundwater flow models. Water Resour. Res. **40**(W01504), 1–11 (2004). doi:http://dx.doi.org/10.1029/2003WR002199
483. Singh, V.: Kinematic Wave Modeling in Water Resources: Surface-Water Hydrology. Wiley, New York (1996)
484. Smith, I., Griffiths, D.: Programming the Finite Element Method, 5th edn. Wiley, Chichester (2004)
485. Smith, I., Farraday, R., O'Connor, B.: Rayleigh-Ritz and Galerkin finite elements for diffusion-convection problems. Water Resour. Res. **9**(3), 593–606 (1973)
486. Sneddon, I.: Elements in Partial Differential Equations. McGraw-Hill, New York (1957)
487. Sonnenveld, P.: CGS: a fast Lanczos-type solver for nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **10**(1), 36–52 (1989)
488. Sposito, G., Chu, S.Y.: The statistical mechanical theory of groundwater flow. Water Resour. Res. **17**(4), 885–892 (1981). doi:http://dx.doi.org/10.1029/WR017i004p00885
489. Springer, J.: Shape-derived anisotropy directions in quadrangle and brick finite elements. Commun. Numer. Methods Eng. **12**(6), 351–357 (1996)
490. Srivastava, R., Yeh, T.C.: Analytical solutions for one-dimensional, transient infiltration toward the water table in homogeneous and layered soils. Water Resour. Res. **27**(5), 753–762 (1991)
491. Steen, P., Aidun, C.: Transition of oscillatory convective heat transfer in a fluid-saturated porous medium. AIAA J. Thermophys. Heat Transf. **1**(3), 268–273 (1987)
492. Strack, O.: Groundwater Mechanics. Prentice Hall, Englewood Cliffs (1989)

493. Strang, G., Fix, G.: An Analysis of the Finite Element Method. Prentice Hall, Englewood Cliffs (1973)
494. Straus, J.: Large amplitude convection in porous media. J. Fluid Mech. **64**(1), 51–63 (1974)
495. Straus, J., Schubert, G.: Three-dimensional convection in a cubic box of fluid-saturated porous material. J. Fluid Mech. **91**(1), 155–165 (1979)
496. Straus, J., Schubert, G.: Modes of finite-amplitude three-dimensional convection in rectangular boxes of fluid-saturated porous material. J. Fluid Mech. **103**, 23–32 (1981)
497. Stüben, K.: Algebraic multigrid (AMG): experiences and comparisons. Appl. Math. Comput. **13**(3–4), 419–451 (1983)
498. Stüben, K.: An introduction to algebraic multigrid. In: Trottenberg, U., Oosterlee, C., Schüller, A. (eds.) Multigrid, pp. 413–532. Elsevier Ltd., Amsterdam (2001)
499. Stüben, K., Clees, T.: SAMG: algebraic multigrid methods for systems, Users's manual. Technical report, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, St. Augustin (2005). http://www.scai.fraunhofer.de/en/business-research-areas/numerical-software/products/samg/samg-user-area.html
500. Stumm, W., Morgan, J.: Aquatic Chemistry. Wiley-Interscience, New York (1981)
501. Suárez, J., Abad, P., Plaza, A., Padrón, M.: Computational aspects of the refinement of 3D tetrahedral meshes. J. Comput. Methods Sci. Eng. **5**(4), 215–224 (2005)
502. Sudicky, E., Unger, A., Lacombe, S.: A noniterative technique for the direct implementation of well bore boundary conditions in three-dimensional heterogeneous formations. Water Resour. Res. **31**(2), 411–415 (1995)
503. Sun, Y., Petersen, J., Clement, T.: Analytical solutions for multiple species reactive transport in multiple dimensions. J. Contam. Hydrol. **35**(4), 429–440 (1999)
504. Sun, Y., Petersen, J., Clement, T., Skeen, R.: Development of analytical solutions for multispecies transport with serial and parallel reactions. Water Resour. Res. **35**(1), 185–190 (1999)
505. Tam, C.: The drag on a cloud of spherical particles in low Reynold number flow. J. Fluid Mech. **38**(3), 537–546 (1969)
506. Tang, D., Frind, E., Sudicky, E.: Contaminant transport in fractured porous media: analytical solution for a single fracture. Water Resour. Res. **17**(3), 555–564 (1981)
507. Taunton, J., Lightfoot, E., Green, T.: Thermohaline instability and salt fingers in a porous medium. Phys. Fluids **15**(5), 748–753 (1972)
508. Taylor, G.: Dispersion of solute matter in solvent flowing slowly through a tube. Proc. R. Soc. Lond. A **219**, 186–203 (1953)
509. Taylor, R.: Solution of linear equations by a profile solver. Eng. Comput. **2**(4), 344–350 (1985)
510. Teza, G., Galgaro, A., De Carli, M.: Long-term performance of an irregular shaped borehole heat exchanger system: analysis of real pattern and regular grid approximation. Geothermics **43**, 45–56 (2012)
511. Theis, C.: The relation between lowering of the piezometric surface and the rate and duration of discharge of a well using ground water storage. Trans. Am. Geophys. Union **16**(2), 519–524 (1935). 16th annual meeting
512. Thiele, K.: Adaptive finite volume discretization of density driven flows in porous media. Ph.D. thesis, Institute of Applied Mathematics, University of Erlangen-Nürnberg, Germany (1999)
513. Thiele, M., Diersch, H.J.: 'Overshooting' effects due to hydrodispersive mixing of saltwater layers in aquifers. Adv. Water Resour. **9**(1), 24–33 (1986)
514. Thompson, J., Soni, B., Weatherill, N.: Handbook of Grid Generation. CRC, Boca Raton (1999)
515. Tien, C.L., Vafai, K.: Convective and radiative heat transfer in porous media. Adv. Appl. Mech. **27**, 225–281 (1989)
516. Toffoli, T., Margolus, N.: Cellular Automata Machines. MIT, Cambridge (1988)
517. Toride, N., Leij, F., van Genuchten, M.: The CXTFIT code for estimating transport parameters from laboratory or field tracer experiments. Vers. 2.1. Technical report No. 137, US Salinity Laboratory, Riverside (1999)

518. Trevisan, O., Bejan, A.: Natural convection with combined heat and mass transfer buoyancy effects in a porous medium. Int. J. Heat Mass Transf. **28**(8), 1597–1611 (1985)
519. Trottenberg, U., Oosterlee, C., Schüller, A.: Multigrid. Elsevier Ltd., Amsterdam (2001)
520. Truesdell, C.: Rational thermodynamics: a course of lectures on selected topics. McGraw-Hill, New York (1969)
521. Truesdell, C., Toupin, R.: Principles of classical mechanics and field theory. In: Flügge, S. (ed.) Handbuch der Physik, vol. III/1, pp. 700–704. Springer, Berlin (1960)
522. Turner, J.: Buoyancy Effects in Fluids. Cambridge University Press, New York (1979)
523. Turner, J.: Multicomponent convection. Annu. Rev. Fluid Mech. **17**, 11–44 (1985)
524. Turner, J.: Laboratory models of double-diffusive processes. In: Brandt, A., Fernando, H. (eds.) Double-Diffusive Convection. Geophysical Monograph, vol. 94, pp. 11–29. American Geophysical Union, Washington, DC (1995)
525. Turner, M., Clough, R., Martin, H., Topp, L.: Stiffness and deflection analysis of complex structures. J. Aerosp. Sci. **23**(9), 805–823 (1956)
526. Tyvand, P.: Thermohaline instability in anisotropic porous media. Water Resour. Res. **16**(2), 325–330 (1980)
527. US Salinity Laboratory: STANMOD (STudio of ANalytical MODels): computer software for evaluating solute transport in porous media using analytical solutions of convection-dispersion equation. Technical report, US Salinity Laboratory, Riverside (2012). http://www.ars.usda.gov/services/software/software.htm
528. Vachaud, G., Vauclin, M.: Comments on 'a numerical model based on coupled one-dimensional Richards and Boussinesq equations' by Mary F. Pikul, Robert L. Street, and Irwin Remson. Water Resour. Res. **11**(3), 506–509 (1975). doi:http://dx.doi.org/10.1029/WR011i003p00506
529. Vachaud, G., Vauclin, M., Khanji, D.: Étude expérimentale des transferts bidimensionnels dans la zone non-saturée. Application á l'étude du drainage d'une nappe á surface libre. Houille Blanche (1), 65–74 (1973)
530. Vadasz, P.: Local and global transitions to chaos and hysteresis in a porous layer heated from below. Transp. Porous Media **37**(2), 213–245 (1999)
531. Vadasz, P., Olek, S.: Computational recovery of the homoclinic orbit in porous media convection. Int. J. Non-Linear Mech. **34**(6), 1071–1075 (1999)
532. Vadasz, P., Olek, S.: Weak turbulence and chaos for low Prandtl number gravity driven convection in porous media. Transp. Porous Media **37**(1), 69–91 (1999)
533. Vadasz, P., Olek, S.: Route to chaos for moderate Prandtl number convection in a porous layer heated from below. Transp. Porous Media **41**(2), 211–239 (2000)
534. Vafai, K.: Handbook of Porous Media, 2nd edn. Taylor and Francis, Boca Raton (2005)
535. van der Vorst, H.: Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems. SIAM J. Sci. Stat. Comput. **13**(2), 631–644 (1992)
536. van Genuchten, M.: Calculating the unsaturated hydraulic conductivity with a new closed form analytical model. Technical report 78-WR-08, Water Resources Program, Princeton University, Princeton (1978)
537. van Genuchten, M.: Mass transport in saturated-unsaturated madia: one-dimensional solutions. Technical report 78-WR-11, Water Resources Program, Princeton University, Princeton (1978)
538. van Genuchten, M.: Numerical solution of the one-dimensional saturated-unsaturated flow equation. Technical report 78-WR-09, Water Resources Program, Princeton University, Princeton (1978)
539. van Genuchten, M.: A closed form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. **44**(5), 892–898 (1980)
540. van Genuchten, M., Alves, W.: Analytical solutions of the one-dimensional convective-dispersive solute transport equation. Technical report technical bulletin number 1661, p. 149, US Department of Agriculture (1982)
541. van Genuchten, M., Gray, W.: Analysis of some dispersion corrected numerical schemes for solution of the transport equation. Int. J. Numer. Methods Eng. **12**(3), 387–404 (1978)

542. van Genuchten, M., Wagenet, R.: Two-site/two-region models for pesticide transport and degradation: theoretical development and analytical solutions. Soil Sci. Soc. Am. J. **53**(5), 1303–1310 (1989)
543. van Reeuwijk, M., Mathias, S., Simmons, C., Ward, J.: Insights from a pseudospectral approach to the Elder problem. Water Resour. Res. **45**(W04416), 1–13 (2009). doi:http://dx.doi.org/10.1029/2008WR007421
544. VDI: VDI-Wärmeatlas: Wärmeübertragung bei der Strömung durch Rohre (heat transfer in flow through pipes). Tech. rep., VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen (2006)
545. Verfürth, R.: A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. Wiley/Teubner, New York/Stuttgart (1996)
546. Verruijt, A.: Theory of Groundwater Flow. MacMillan, London (1970)
547. Vinsome, P.: ORTHOMIN, an iterative method for solving sparse sets of simultaneous linear equations, paper SPE 5739. In: Proceedings of the 4th Symposium on Reservoir Simulation, Society of Petroleum Engineers of AIME, Los Angeles, pp. 149–159 (1976)
548. Volker, R., Rushton, K.: An assessment of the importance of some parameters for seawater intrusion and a comparison of dispersive and sharp-interface modeling approaches. J. Hydrol. **56**(3–4), 239–250 (1982)
549. Volocchi, A., Street, R., Roberts, P.: Transport of ion-exchanging solutes in groundwater: chromatographic theory and field simulation. Water Resour. Res. **17**(5), 1517–1527 (1981)
550. Voss, C.: A finite-element simulation model for saturated-unsaturated fluid-density-dependent ground-water flow with energy transport or chemically-reactive single-species solute transport. Technical report, Water Resources Investigations, report 84-4369, p. 409, US Geological Survey (1984)
551. Voss, C.: USGS SUTRA code – history, practical use, and application in Hawaii. In: Bear, J., et al. (eds.) Seawater Intrusion in Coastal Aquifers: Concepts, Methods and Practices, pp. 249–313. Kluwer Academic, Dordrecht (1999)
552. Voss, C., Souza, W.: Variable density flow and solute transport simulation of regional aquifers containing a narrow freshwater-saltwater transition zone. Water Resour. Res. **23**(10), 1851–1866 (1987)
553. Šimůnek, J., Vogel, T., van Genuchten, M.: The SWMS-2D code for simulating water flow and solute transport in two-dimensional variably saturated media. Technical report no. 126, US Salinity Laboratory, Riverside (1992)
554. Šimůnek, J., Kodešová, R., Gribb, M., van Genuchten, M.: Estimating hysteresis in the soil water retention function from cone permeameter experiments. Water Resour. Res. **35**(5), 1329–1345 (1999)
555. Wait, R., Mitchell, A.: Finite Element Analysis and Applications. Wiley, New York (1985)
556. Walker, K., Homsy, G.: A note on convective instabilities in Boussinesq fluids and porous media. ASME J. Heat Transf. **99**(2), 338–339 (1977)
557. Watson, D.: Computing n-dimensional Delaunay tesselation with application to Voronoi polytopes. Comput. J. **24**(2), 167–172 (1981)
558. Webb, S.: Generalization of Ross' tilted capillary barrier diversion formula for different two-phase characteristic curves. Water Resour. Res. **33**(8), 1855–1859 (1997)
559. Wendland, E.: Numerische Simulation von Strömung und hochadvektivem Stofftransport in geklüftetem, porösem Medium (numerical simulation of flow and advection-dominant mass transport in fractured and porous medium). Ph.D. thesis, Dissertation, Ruhr University Bochum, Bochum, Germany (1996). Report No. 96-6
560. Wendland, E., Himmelsbach, T.: Transport simulation with stochastic aperture for a single fracture – comparison with a laboratory experiment. Adv. Water Resour. **25**(1), 19–32 (2002)
561. Werner, A., Bakker, M., Post, V., Vandenbohede, A., Lu, C., Ataie-Ashtiani, B., Simmons, C., Barry, D.: Seawater intrusion processes, investigation and management: recent advances and future challenges. Adv. Water Resour. **51**, 3–26 (2013)
562. Whitaker, S.: The Forchheimer equation: a theoretical development. Transp. Porous Media **25**(1), 27–61 (1996)

563. Whitaker, S.: The Method of Volume Averaging. Kluwer Academic, Dordrecht (1999)
564. Wiedemeier, T., Swanson, M., Moutoux, D., Kinzie Gordon, E., Wilson, J., Wilson, B., Kampbell, D., Haas, P., Miller, R., Hansen, J., Chapelle, F.: Technical protocol for evaluating natural attenuation of chlorinated solvents in ground water. Technical report EPA/600/R-98/128, US Environmental Protection Agency (1998)
565. Wilkinson, F.: Chemical Kinetics and Reaction Mechanisms. Van Nostrand Reinhold, New York (1980)
566. Williams, G., Miller, C., Kelley, C.: Transformation approaches for simulating flow in variably saturated porous media. Water Resour. Res. **36**(4), 923–934 (2000)
567. Wolfram, S.: Cellular Automata and Complexity. Addison-Wesley, Reading (1994)
568. Wood, W., Lewis, R.: A comparison of time marching schemes for the transient heat conduction equation. Int. J. Numer. Methods Eng. **9**(3), 679–689 (1975)
569. Wooding, R.: Steady state free thermal convection of liquid in a saturated permeable medium. J. Fluid Mech. **2**(3), 273–285 (1957)
570. Wooding, R.: Variable-density saturated flow with modified Darcy's law: the salt lake problem and circulation. Water Resour. Res. **43**(W02429), 1–10 (2007). doi:http://dx.doi.org/10.1029/2005WR004377
571. Wooding, R., Tyler, S., White, I.: Convection in groundwater below an evaporating salt lake, 1. Onset of instability. Water Resour. Res. **33**(6), 1199–1217 (1997)
572. Wooding, R., Tyler, S., White, I., Anderson, P.: Convection in groundwater below an evaporating salt lake, 2. Evolution of fingers or plumes. Water Resour. Res. **33**(6), 1219–1228 (1997)
573. Woods, J., Carey, G.: Upwelling and downwelling behavior in the Elder-Voss-Souza benchmark. Water Resour. Res. **43**(W12405), 1–12 (2007). doi:http://dx.doi.org/10.1029/2006WR004918
574. Xie, Y., Simmons, C., Werner, A., Ward, J.: Effect of transient solute loading on free convection in porous media. Water Resour. Res. **46**(W11511), 1–16 (2010). doi:http://dx.doi.org/10.1029/2010WR009314
575. Xie, Y., Simmons, C., Werner, A.: Speed of free convective fingering in porous media. Water Resour. Res. **47**(W11501), 1–16 (2011). doi:http://dx.doi.org/10.1029/2011WR010555
576. Xie, Y., Simmons, C., Werner, A., Diersch, H.J.: Prediction and uncertainty of free convection phenomena in porous media. Water Resour. Res. **48**(2,W02535), 1–12 (2012). doi:http://dx.doi.org/10.1029/2011WR011346
577. Yavuzturk, C., Spitler, J., Rees, S.: A transient two-dimensional finite volume model for the simulation of vertical U-tube ground heat exchangers. ASHRAE Trans. **105**(2), 465–474 (1999)
578. Yeh, G.T.: On the computation of Darcy velocity and mass balance in the finite element modeling of groundwater flow. Water Resour. Res. **17**(5), 1529–1534 (1981)
579. Yeh, G.T.: Computational Subsurface Hydrology: Fluid Flows. Kluwer Academic, Dordrecht (1999)
580. Yeh, G.T.: Computational Subsurface Hydrology: Reactions, Transport, and Fate. Kluwer Academic, Dordrecht (2000)
581. Yeh, G.T., Chen, J.R., Bensabat, J.: A three-dimensional finite-element model of transient free surface flow in aquifers. In: Peters, A., et al. (eds.) Proceedings of the 10th International Conference on Computational Methods in Water Resources, Heidelberg, vol. 1, pp. 131–138. Kluwer Academic, Dordrecht (1994)
582. Younes, A., Ackerer, P.: Empirical versus time stepping with embedded error control for density-driven flow in porous media. Water Resour. Res. **46**(W08523), 1–8 (2010). doi:http://dx.doi.org/10.1029/2009WR008229
583. Younes, A., Ackerer, P., Mosé, R.: Modeling variable density flow and solute transport in porous medium: 2. Re-evaluation of the salt dome flow problem. Transp. Porous Media **35**(3), 375–394 (1999)
584. Yu, C.C., Heinrich, J.: Petrov-Galerkin methods for the time-dependent convective transport equation. Int. J. Numer. Methods Eng. **23**(5), 883–901 (1986)

585. Yu, C.C., Heinrich, J.: Petrov-Galerkin methods for multidimensional, time-dependent, convective-diffusion equations. Int. J. Numer. Methods Eng. **24**(11), 2201–2215 (1987)
586. Zgainski, F.X., Coulomb, J.L., Maréchal, Y., Claeyssen, F., Brunotte, X.: A new family of finite elements: the pyramidal elements. IEEE Trans. Magn. **32**(3), 1393–1396 (1996)
587. Zheng, C., Bennett, G.: Applied Contaminant Transport Modeling: Theory and Practice. Van Nostrand Reinhold, New York (1995)
588. Zidane, A., Younes, A., Huggenberger, P., Zechner, E.: The Henry semianalytical problem for saltwater intrusion with reduced dispersion. Water Resour. Res. **48**(W06533), 1–10 (2012). doi:http://dx.doi.org/10.1029/2011WR011157
589. Zienkiewicz, O., Cheung, Y.: The Finite Element Method in Structural and Continuums Mechanics. McGraw-Hill, London (1967)
590. Zienkiewicz, O., Taylor, R.: The Finite Element Method. Volume 1: The Basis, 5th edn. Butterworth-Heinemann, Oxford (2000)
591. Zienkiewicz, O., Taylor, R.: The Finite Element Method. Volume 2: Solid and Structural Mechanics, 5th edn. Butterworth-Heinemann, Oxford (2000)
592. Zienkiewicz, O., Taylor, R.: The Finite Element Method. Volume 3: Fluid Dynamics, 5th edn. Butterworth-Heinemann, Oxford (2002)
593. Zienkiewicz, O., Zhu, J.: A simple error estimator and adaptive procedure for practical engineering analysis. Int. J. Numer. Methods Eng. **24**(2), 337–357 (1987)
594. Zienkiewicz, O., Zhu, J.: The superconvergent patch recovery and *a posteriori* error estimates. Part 1: The recovery technique. Int. J. Numer. Methods Eng. **33**(7), 1331–1364 (1992)
595. Zienkiewicz, O., Heinrich, J., Huyakorn, P., Mitchell, A.: An ‘upwind’ finite element scheme for two-dimensional convective transport equations. Int. J. Numer. Methods Eng. **11**(1), 131–143 (1977)

Index

H.-J.G. Diersch, *FEFLOW*, DOI 10.1007/978-3-642-38739-5,