TOUGH: MODEL USE, CALIBRATION, AND VALIDATION

S. Finsterle*, M. B. Kowalsky*, and K. Pruess*

ABSTRACT. *Simulation models for subsurface flow and transport require the estimation of many disparate parameters to capture properties and processes that cannot be explicitly represented in the model. The main challenges for model calibration reside in the formulation of the forward model, the identification of influential parameters, the selection of appropriate calibration data with sufficient information content about these parameters, and the performance of a robust and efficient minimization algorithm capable of handling complex topologies of the resulting objective function. The large variety of subsurface systems and related engineering problems that are addressed by numerical simulation prevents us from compiling definite lists of critical input parameters and desirable observation data to be used for model calibration. Instead, we describe the approach that we consider necessary to arrive at a defensible model for predictive simulations. This process includes sensitivity analyses and synthetic inversions prior to data collection, and detailed residual, error, and uncertainty analyses after automatic model calibration. We illustrate the approach by discussing applications of the iTOUGH2 simulation-optimization code, which supports the development of TOUGH2 models for nonisothermal, multiphase, multicomponent flow and transport processes in fractured porous geological media. We conclude that the calibration of models representing complex subsurface systems remains challenging and requires research that is targeted at the needs of scientists and engineers who address simulation problems of practical relevance.*

Keywords. *iTOUGH2, Nonisothermal multiphase flow simulation, Parameter estimation, Sensitivity analysis, TOUGH2, Uncertainty quantification.*

The TOUGH codes simulate nonisothermal flow of multiphase, multicomponent fluids in porous and fractured geologic media (Pruess, 1991; Pruess et al., 1999; http://esd.lbl.gov/TOUGH2). Developed at the Lawrence Berkeley National Laboratory in the early 1980s primarily for geothermal reservoir engineering, the suite of simulators (table 1) is now widely used by universities, government organizations, and private industry for applications related to nuclear waste disposal, environmental remediation problems, geothermal energy production, oil and gas reservoirs, gas hydrate deposits, geological carbon sequestration, vadose zone hydrology, and other uses that involve coupled thermal, hydrological, geochemical, and geomechanical processes in permeable media (Pruess, 2004; Finsterle et al., 2008). The TOUGH suite of simulators is continually updated, with new equation-of-state (EOS) modules being developed, and refined process descriptions implemented into its framework. Notably, EOS property modules for mixtures of water, NaCl, and sub- and supercritical CO_2 have been developed (Pruess, 2005, 2011) and are widely used for the analysis of geologic carbon sequestration processes.

TOUGH+ is the result of a re-engineering effort in which the capabilities of TOUGH2 have been recast in a modular design that adheres to the tenets of object-oriented programming as implemented in Fortran 95. TOUGH+ also includes new capabilities. In particular, the range of thermodynamic properties of water has been extended to enable modeling of freezing and thawing in permafrost regions and, specifically, for the simulation of processes in hydrate-bearing formations, which involve formation and dissociation of ice-like hydrates with associated changes in porosity, permeability, and two-phase flow characteristics (Moridis et al., 2008; Kowalsky et al., 2010).

TOUGHREACT adds transport of chemically reactive species to the nonisothermal multiphase flow capabilities of TOUGH2 (Xu and Pruess, 2001; Xu et al., 2004). Interactions between mineral assemblages and fluids can occur under local equilibrium or with kinetic rates. The gas phase can be chemically active. Precipitation and dissolution reactions can change formation porosity and permeability, which in turn may modify the two-phase flow properties of the rock.

The iTOUGH2 simulation-optimization code provides inverse modeling capabilities as well as formal sensitivity and uncertainty propagation analyses for most TOUGH modules (Finsterle, 1992, 1994, 1997a, 1997b, 1997c; http://esd.lbl.gov/iTOUGH2). The iTOUGH2 optimization and analysis methods can also be utilized for the automatic calibration of any standalone simulation program that uses ASCII input and output files (Finsterle and Zhang, 2011a). iTOUGH2 is continually upgraded to incorporate new developments of the TOUGH2 forward simulators, as well as to update its optimization algorithms and various analysis tools. For example, iTOUGH2 has been expanded to jointly invert hydrogeological and geophysical data for high-resolution characterization of the subsurface (Kowalsky et al., 2004, 2005,

***The authors are **Stefan Finsterle,** Staff Scientist, **Michael B. Kowalsky,** Research Scientist, and **Karsten Pruess,** Senior Scientist Emeritus, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California. **Corresponding author:** Stefan Finsterle, Earth Sciences Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, MS 90-1116, Berkeley, CA 94720; phone: 510-486-5205; e-mail: SAFinsterle@lbl.gov.

2008, 2011; Finsterle and Kowalsky, 2008).

The TOUGH codes have a broad developer and user community, as is reflected by the large international attendance at the TOUGH symposia, workshops, and training courses. Several special issues of peer-reviewed journals have been dedicated to developments and applications of the TOUGH simulators. Links to these articles in special issues, along with an extensive bibliography of journal articles, manuals, proceedings, and reports, as well as announcements of symposia and short courses, can be found at http://esd.lbl.gov/TOUGH2, which also contains information about licensing of the source codes, copyrighted by the University of California.

Table 1. Overview of TOUGH simulators.

Simulator	Phases, Components, and Processes[a]	Released	Key References[b]
MULKOM	Research code for nonisothermal multiphase, multicomponent flows of Newtonian and non-Newtonian fluids	No public release	Pruess (1983)
TOUGH	Nonisothermal flow of water and air in aqueous and gaseous phase	1987	Pruess (1987)
TOUGH2	Nonisothermal flow of water and NCG in aqueous and gaseous phase	1991	Pruess (1991)
iTOUGH	Inverse modeling for TOUGH	No public release	Finsterle (1992)
T2VOC	Nonisothermal flow of water, air, and VOCs in aqueous, gaseous, and NAPL phase for environmental applications	1995	Falta et al. (1995)
iTOUGH2	Inverse modeling, sensitivity analysis, and uncertainty propagation analysis for TOUGH2 (Pruess, 1991) and additional EOS modules	1997	Finsterle (1997a, 1997b, 1997c)
TOUGH2 V2	Nonisothermal multiphase, multicomponent flow	1999	Pruess et al. (1999)
TMVOC	Nonisothermal flow of water, air, multiple VOCs, and NCGs in aqueous, gaseous, and NAPL phase	2002	Pruess and Battistelli (2002)
TOUGHREACT	Nonisothermal multiphase flow and reactive transport including equilibrium and kinetic mineral dissolution and precipitation, chemically active gases, intra-aqueous and sorption reaction kinetics, and biodegradation	2004	Xu and Pruess (2001), Xu et al. (2004)
TOUGH+	Re-engineered and expanded version of TOUGH2 simulator, specifically for the simulation of hydrate-bearing geologic media	2008	Moridis (2003), Moridis et al. (2008)
TOUGH-FLAC	Research code for coupled multiphase flow and thermal-geomechanical processes; links TOUGH2 and FLAC3D (Itasca, 1997)	No public release	Rutqvist et al. (2002)
TOUGH-MP	Massively parallel version of TOUGH2	2008	Zhang et al. (2008)
ECO2M	Multiphase flow of sub- and supercritical CO_2	2011	Pruess (2011)

[a] NAPL = non-aqueous phase liquid, NCG = noncondensible gas, and VOC = volatile organic compound.
[b] For a complete list, see http://esd.lbl.gov/TOUGH+.

TOUGH DESCRIPTION

SPATIAL AND TEMPORAL SCALES AND DISCRETIZATION

TOUGH2 is a general-purpose numerical simulator that solves coupled mass- and heat-balance equations governing the nonisothermal flow of multiphase, multicomponent fluids in permeable media. The governing equations are solved using the integral finite difference method (IFDM; Edwards, 1972; Narasimhan and Witherspoon, 1976). In the IFDM, local fluxes are evaluated in one dimension; the method avoids any reference to a global coordinate system and is locally mass conservative. The flexibility offered by this discretization scheme allows the user to effectively represent geological systems with, for example, simple one-dimensional systems or complex three-dimensional geometries that may include engineered structures or irregularly shaped geologic units. Multiscale features (such as fractures embedded in the matrix) can be handled using multicontinua approaches, which are implemented through simple preprocessing of geometric data (Doughty, 1999). Similarly, radial, spherical, and other non-Cartesian flow regimes are naturally accommodated by the mesh geometry, obliterating the need to reformulate the underlying flow and transport equations.

While the governing equations are formulated on the continuum scale, there is no inherent limitation on the spatial scale to which the code can be applied. TOUGH2 has been used to simulate pore-scale processes, to analyze data from cores and small-scale laboratory experiments, to design field tests, to examine impacts on the scale of engineered structures, to predict reservoir behavior, and to assess the response to natural changes and human interventions on the basin scale.

The range in spatial scales is mirrored by the range in temporal scales. TOUGH2 analyses span periods from short-term responses during well testing, to the projected operating time of a reservoir, to the assessment of long-term safety of geologic storage of CO_2 and radionuclides, to the analysis of geochemical processes occurring over geologic times. The differences in time constants inherent in a variety of coupled processes (e.g., flow, transport, and reaction) and subsystems (e.g., wellbore and reservoir flow, surface water and groundwater flow) lead to computational challenges. Transient effects requiring short time steps can result from strongly time-dependent forcing terms imposed at internal and external boundaries, from system heterogeneity encountered by a pressure, saturation, or reaction front, or from abrupt changes of the state as the system evolves (e.g., appearance or disappearance of a phase). These challenges are partly addressed by a

flexible time-stepping scheme, and by a fully coupled, implicit solution of all governing equations.

Heterogeneity on multiple scales is prevalent in all natural subsurface environments, with property values potentially changing over many orders of magnitude. Whether such heterogeneity needs to be represented explicitly or can be lumped into effective parameters strongly depends on the processes being considered as well as the spatial and temporal scale of interest. The effective property values estimated by inverse modeling are therefore directly linked to the chosen parameterization of heterogeneity, as discussed below.

On the scale of an individual grid block, the description of thermodynamic conditions is based on the assumption of local equilibrium of all phases. The validity of this local equilibrium assumption again depends on the spatial and temporal scale of interest. This means that a violation of the assumption can be addressed (within practical constraints) by refining the computational grid.

In TOUGH2, time is discretized fully implicitly as a first-order backward finite difference, an approach that, together with full upstream weighting of flux terms at interfaces, achieves unconditional stability and avoids impractical time step limitations in flow problems involving phase (dis-)appearances. This scheme comes at the expense of increased numerical dispersion, an issue that can be addressed as discussed by Oldenburg and Pruess (2000).

In multiphase systems, fluid and formation properties are usually nonlinear functions of the primary thermodynamic variables. The spatial and temporal discretization of the governing equations results in a set of strongly coupled nonlinear algebraic equations representing mass and energy residuals. The time-dependent primary thermodynamic variables in all grid blocks are the unknowns. The governing equations in their discretized form are solved simultaneously using Newton-Raphson iterations, which requires setting up a Jacobian matrix that holds the partial derivatives of the residuals with respect to the primary variables. Within each Newton-Raphson iteration, the set of linear equations is solved using a preconditioned conjugate gradient solver. Details about the numerical solution scheme implemented in TOUGH2 can be found in Pruess et al. (1999) and Moridis and Pruess (1998).

PHASES, COMPONENTS, AND PROCESSES

The different TOUGH modules numerically solve mass-balance equations for select chemical components (or pseudo-components, such as air). All modules contain water, and may include non-condensible gases, solutes, volatile organic compounds, and radionuclides. These components are transported through the geologic media in one, two, or three fluid phases (gaseous, aqueous, and non-aqueous phase liquids). Some components may also precipitate and dissolve or adsorb and desorb, i.e., become or interact with an immobile solid phase. The partitioning of each component in each phase and the appearance or disappearance of a phase is determined by the local thermodynamic conditions, which are dynamically calculated for each grid block. For example, liquid water may vaporize as temperature rises or dry gas is injected, increasing the salinity of the disappearing liquid phase, which potentially leads to the precipitation of NaCl. Phases may completely vanish (e.g., by vaporization of the liquid or compression and dissolution of the gas phase) or newly evolve (e.g., by condensation of vapor or degassing of air dissolved in a liquid). The number of phases determines which state variables are independent; consequently, the (dis)appearance of a phase requires dynamic switching of primary thermodynamic variables in each grid block.

The main processes considered in the TOUGH simulators are advective fluid flow described with a multiphase extension of Darcy's law, which includes relative permeability and capillary pressure effects; in addition, there is diffusive mass transport in all phases. Heat flow occurs by conduction and convection and includes sensible and latent heat effects. Many refinements and enhancements have been made to this basic process description, including non-Darcy and overland flow, non-Newtonian fluids and Klinkenberg effects, colloidal transport and radioactive decay, hysteresis and dispersion, as well as water uptake by roots. In particular, the TOUGHREACT code includes geochemical and microbiological reactions that are coupled to flow and transport processes. TOUGHREACT considers a variety of subsurface thermo-physical-chemical processes for a wide range of hydrological and chemical conditions. Interactions between mineral assemblages and fluids can be modeled assuming local equilibrium or kinetic rate laws. The gas phase may be chemically active, and precipitation and dissolution reactions may change formation porosity and permeability. Reactions among primary species (including intra-aqueous and sorption reaction kinetics and biodegradation) are described using a general rate law that accounts for multiple mechanisms and multiple products, concentration-dependent (Monod) rate expressions, and inhibition terms. The reader is referred to the various user's guides (listed in table 1 and downloadable at http://esd.lbl.gov/TOUGH2) for detailed descriptions of the physical processes, user features, code usage, and application sample problems.

The iTOUGH2 simulation-optimization code provides inverse modeling, sensitivity analysis, and uncertainty quantification for all TOUGH modules mentioned above; these modules are either fully integrated into the iTOUGH2 framework or can be accessed through a flexible interface. The following discussion focuses on the iTOUGH2 calibration approach as applied to subsurface nonisothermal multiphase flow problems.

TOUGH CALIBRATION AND UNCERTAINTY ANALYSIS

CODE VERIFICATION VS. MODEL CALIBRATION

As discussed above, TOUGH is a physics-based general-purpose simulator capable of modeling a large variety of processes and systems. It is therefore essential to distinguish between the computer code, which is general, and the model, which is site and problem specific. While the code can be tested to make sure the governing equations are solved correctly (i.e., within numerical errors that are deemed acceptable), calibration is performed for a particular model. Moreover, calibration standards and acceptance criteria for testing the model's ability to make meaningful predictions depend on the overall purpose of the model. For example, if a model is developed solely to examine the relative influence of parameters that are linearly related to a performance measure of interest, no calibration is needed. However, if the same linear model is to be used for predictive purposes, the values of parameters identified as influential may need to be estimated by calibration. Moreover, if the model is highly nonlinear, a (local) sensitivity analysis depends on the parameter set, potentially requiring a calibration step prior to the sensitivity analysis; alternatively, a global sensitivity analysis can be performed. If the objective of a model is to predict a steady-state flow field, there may be no need to estimate porosity, as its impact on simulation results such as flow rates and saturations may be insignificant for a given problem. Nevertheless, porosity may need to be estimated if the highly influential permeability value is determined in a calibration that uses transient data, which usually leads to a (statistical) correlation between permeability and porosity. In such a case, if there is an error in the assumed porosity value, the estimated permeability may be biased. As a final example, it is obvious that the strategy followed to test the ability of the model to make predictions depends on its intended use. The most sensible modeling approach to be chosen depends on whether we are interested in the average behavior or the tail of a contaminant plume, as this affects the associated risk distribution. Moreover, the way uncertainties are represented, and which performance measures are used to assess our confidence in the model's ability to make an accurate, robust, and defensible prediction also depend on the intended use of the model.

In summary, no general rules can, or should, be provided about which parameters require calibration, which type of data should be collected for calibration of the model and for testing its ability to make reasonable predictions, and which specific modeling, inversion, optimization, or uncertainty quantification approach should be used. These questions have to be answered for each case individually. In fact, the design of a modeling study should proceed backwards, starting at the end with a clear understanding of the ultimate goal of the project and the role that modeling will play. Once the model objectives are formulated, performance measures can be selected, along with the acceptable level of uncertainty with which they need to be predicted by the model. Next, the required model complexity needs to be determined, before the model is suitably parameterized and the impact of these parameters on the performance measures can be identified. This analysis in turn determines the acceptable level of uncertainty with which the most influential parameters need to be estimated. If inverse modeling is used for parameter estimation, the data collection strategy (which includes selection of data type, design of the monitoring network, optimization of sampling time and frequency, determination of required measurement accuracy, etc.) directly depends on the acceptable estimation uncertainty. Note that all these modeling evaluations (i.e., sensitivity, uncertainty propagation, and synthetic inverse analyses) can be performed before calibration data are collected and can support experimental design. Once the data are available, the analysis is executed in reverse order. Characterization data with high information content are collected, and parameters are determined by inverse modeling. The uncertainty of the estimated parameters is assessed and propagated through the prediction model. Finally, performance measures are evaluated for ultimate decision support.

The following subsections describe the parameters that are typically estimated by inverting a multiphase flow and transport model, and the observations that are used as calibration points. We also describe the features of the iTOUGH2 simulation-optimization code that support this analysis. While most of the concepts described here have been discussed in the seminal paper by Carrera and Neuman (1986) and many other scientific articles and textbooks, we mainly refer to publications that demonstrate their application to multiphase inverse problems solved by iTOUGH2.

CALIBRATION PARAMETERS

We define a parameter as a real-valued number that is associated with one or multiple input variables of the TOUGH simulator or one of its pre- or post-processors. In the context of sensitivity analysis, parameter estimation, and uncertainty quantification, parameters are a subset of those TOUGH inputs that are considered unknown or uncertain and whose values are thus varied. The term parameterization refers to the process of defining all parameters that are potentially subjected to these analyses.

Unlike *ab initio* models, which are derived from first principles at the level of established physical laws, most environmental models are based on or include empirical submodels that capture simplifying assumptions, lumped features, and upscaled processes. While some of these empirical submodels closely reflect physical processes on the continuum scale (Darcy's and Fick's laws being prominent examples), others are more abstract or entirely removed from physical meaning, such as purely data-driven, transfer-function based models. The coefficients of these submodels are the fitting parameters potentially estimated by inverse modeling. This means that the formulation of the conceptual model, which includes the

governing equations implemented in the numerical code, predetermines a significant part of the model's parameterization. In addition, the definition of geometric objects that represent site-specific hydrogeologic features is also part of the parameterization process. These features may be described deterministically or stochastically. In particular, the location of a river may be known, and its hydraulic interaction with the aquifer may be parameterized using a small number of riverbed leakage coefficients. On the other hand, the subsurface is inherently heterogeneous, which poses a significant challenge for model calibration, as heterogeneity results in a high-dimensional inverse problem if estimation of separate property values is attempted for each computational cell of the model. However, heterogeneity, which often exhibits structural and random components, may be represented using various geostatistical methods, which allows the modeler to drastically reduce the number of parameters that need to be estimated (e.g., requiring only the geostatistical parameters and property values at select conditioning points to be estimated; see Finsterle and Kowalsky, 2008, and Jung et al., 2011, for examples using iTOUGH2). Other approaches (such as the use of facies or depositional models, or a discrete fracture network generator) are also capable of flexibly generating realizations of subsurface heterogeneity using a limited number of parameters that can be estimated by model calibration.

The above discussion highlights the importance of parameterization, which is an essential step in the development of any predictive model. Parameterization is an even more critical step when performing inverse modeling, as it affects the data needs, the inversion approach, and ultimately the success of model calibration. Note that the estimated parameter values are strictly valid only for the given model structure, which includes parameterization and the decision of which coefficients are considered unknown or uncertain and which are fixed during the inversion. In other words, estimated parameters are always model-related, process-specific, and scale-dependent, with implications for their validity in prediction models.

The question arises about the optimal level of parameterization, and how many parameters can be concurrently estimated during model calibration. It is obvious that the number and type of parameters that can be estimated depend on the amount, type, and quality of available calibration data. The fit to a given data set can always be improved if more parameters are added, at the expense of increased estimation uncertainty and the risk of overfitting, where information about the parameters is incorrectly extracted from random data errors. The size of the solution space (or, complementarily, the calibration null space) can be determined dynamically during model calibration based on sensitivity coefficients, correlation measures, or the properties of the information matrix (Tonkin and Doherty, 2005; Moore and Doherty, 2006). Moreover, regularization, in which model parsimony is appropriately weighted relative to a data misfit criterion, can help avoid these pitfalls. In addition to these formal approaches, it is essential that the modeler has an excellent understanding of the processes being considered and the features of the system that are likely to be relevant in order to arrive at a defensible parameterization.

As mentioned above, definitive statements about which parameters need to be estimated by model calibration cannot be made. Nevertheless, some guidance can be given. For example, for a physics-based simulator, any fundamental parameter directly linked to a physical law does not need to be estimated. This is the reason why fluid properties for a wide range of thermodynamic conditions are internally provided in the TOUGH simulators, i.e., they are not user-specified input parameters that can be subjected to parameter estimation. An exception are rheologic properties of engineered fluids that have unknown reactions with *in situ* fluids (for an example, see Gallagher and Finsterle, 2004).

Formation parameters that may not need to be estimated include those that can be readily measured, that tend to vary over a relatively small range, and that have limited influence on only one aspect of a coupled process. An example of such a parameter is thermal conductivity for fully saturated porous media. However, thermal conductivity under two-phase flow conditions is difficult to measure, and conductive heat losses in a laboratory setting may have a significant impact on the results of the experiment. Moreover, conduction may be the dominant heat transfer process under certain conditions. In such cases, thermal conductivity needs to be determined for the specific model, i.e., it cannot be taken from literature values; examples are discussed by Engelhardt and Finsterle (2003) and Freifeld et al. (2008).

Observations of a hydrogeological system typically reflect its response to fluid flow. Consequently, the absolute permeability is generally a highly influential parameter. Permeability is a property of the porous medium (i.e., it does not include fluid properties, as hydraulic conductivity does); it is an effective parameter that may be scale-dependent as a result of unresolved heterogeneity; and it may be anisotropic as a result of the sub-grid-block structure of this heterogeneity. Since permeability is usually assumed constant on the scale of an individual cell of the computational grid, or even within a predefined larger zone that may represent a hydrostratigraphic unit, permeability is obviously a model-related parameter. This means that the calibrated permeability value depends on the chosen parameterization, and it may be biased if there is an error in the model structure. For example, if the number and geometry of the model layers do not properly represent the actual hydrostratigraphy of the geological formation, the permeability estimates for these zones will be biased in an attempt to, partly, compensate for the structural error. One approach to address this issue is to parameterize and estimate the model structure along with the property values. In particular, geophysical data contain high-resolution information of the subsurface structure and can thus complement hydrogeological data in a joint inversion framework to determine heterogeneous permeability fields. Examples of joint hydrogeophysical inversions using iTOUGH2 are discussed by Kowalsky et al. (2004, 2005, 2008, 2011), Finsterle and Kowalsky (2008), and Lehikoinen et al. (2010).

A main feature of the TOUGH simulators is that they handle multiphase flow processes, where interferences among the aqueous, gas, and oil phases are captured by means of relative permeability and capillary pressure curves. These characteristic curves are implemented as highly nonlinear, parameterized functions. Similar to absolute permeability, the parame-

ters of the relative permeability functions tend to be highly influential on the flow behavior, and thus affect the measurable quantities (such as pressure and saturation) that may be used for model calibration. Some of these parameters (e.g., the residual saturation of the nonwetting phase) are properties that affect key mechanisms of interest, such as the trapping of non-aqueous phase liquids, which is relevant for environmental clean-up, or the trapping of supercritical CO_2 injected for geologic sequestration (Zhang et al., 2011). The parameters of the relative permeability functions can be estimated from pedotransfer functions, from measurements of the water retention curve and application of a suitable pore-connectivity model, or by directly fitting the function to data from two-phase flow experiments. However, each of these approaches has its own challenges and may yield parameter values that are not necessarily applicable for use in a field-scale numerical model. In many applications, the relative permeability near full saturation of the wetting phase has the most significant impact on the system behavior, and if the chosen relative permeability model is oversimplified (e.g., it assumes a unimodal pore-size distribution), these parameters may be inappropriate to predict two-phase flow in a soil that contains macropores. In such cases, it may be better to determine the relative permeability parameters by inverse modeling using data collected on the appropriate scale and under the relevant saturation conditions, despite the fact that the resulting relative permeability functions may not match the laboratory data well over the entire saturation range. Again, the parameters estimated by inverse modeling are process-specific, scale-dependent, and model-related, which is a limitation but also an advantage; these issues are further discussed by Finsterle and Faybishenko (1999).

Similarly, the parameters of the capillary pressure function can be considered effective parameters that should be determined with a particular application and process in mind. The gas-entry pressure, the slope of the curve in the mid-saturation range, and its steepness as it approaches residual saturation are usually related to specific systems or processes that should be emphasized during calibration. Examples of iTOUGH2 calibrations of capillary pressure parameters are given by Finsterle (2000) and Ghezzehei et al. (2004).

Finally, model calibration may not be restricted to the determination of material properties. In fact, initial and boundary conditions as well as sink and source terms are among the most influential factors that may also be highly uncertain. For example, the calibration of a transport model using concentration data will yield unreasonable results unless the location and strength of the contaminant source is either well known or estimated along with the flow and transport parameters. Similarly, the initial temperature and pressure distribution as well as the heat upflow in a geothermal reservoir need to be determined before production data are used for the calibration of parameters affecting fluid and heat flow (Kiryukhin et al., 2008). Even slight changes in the magnitude and orientation of the groundwater gradient may adversely affect the interpretation of tracer test data. Consequently, such changes may need to be included in the inversion (Kowalsky et al., in review).

In summary, a large variety of parameters may be considered unknown or uncertain. If they are expected to significantly influence the system behavior of interest, or are correlated to other parameters that are, they need to be determined. It is preferable that these parameters be determined concurrently in a joint inverse modeling framework to yield effective parameters that have a reduced risk of being biased due to errors in related parameters, and that have estimation uncertainties that reflect this correlation structure. These parameters refer to hydraulic, thermal, geochemical, biological, and mechanical material properties but may also include the system state (i.e., initial and boundary conditions), geometrical parameters (such as fracture spacing, and location and size of a source domain), geostatistical properties (e.g., variance and correlation length of a heterogeneous property field), or petrophyscial parameters (e.g., when performing hydrogeophysical inversions). The parameters to be analyzed by iTOUGH2 need to be mapped to one or several input variables of the forward model. Moreover, they may be tied to each other and may need to be transformed (e.g., by taking their logarithms) before being assembled in a vector $\mathbf{p}$ of length n. Whether these parameters can be determined with acceptably low estimation uncertainty depends on the availability of complementary, sensitive calibration data of good quality, as will be discussed next.

DATA NEEDS FOR CALIBRATION

When calibrating a model, information about the parameters of interest is extracted from observations of the system response at discrete points in space and time. The measured data only contain information about a parameter if the corresponding model output is sufficiently sensitive to a change in the parameter value. For example, pressure data observed during a well test performed in a saturated aquifer are sensitive to the formation's hydraulic diffusivity, whereas pressure might not be sensitive to any of the two-phase parameters (e.g., the air-entry pressure, residual saturation, or pore-size distribution index) unless a noncondensible gas is injected to create two-phase conditions in the aquifer.

This simple example illustrates that questions about (1) optimal test design (e.g., injection of liquid water vs. injection of a noncondensible gas), (2) the choice of data to be collected (e.g., pressure, saturation, flow rate), and (3) the selection of calibration parameters (e.g., permeability or two-phase flow properties) must be linked to the intended use of the model. Consequently, no generally applicable statements about data needs can be made, as these needs are always related to the objectives of a particular study. Nevertheless, a few principles or guidelines can be given.

First, due to the fact that estimated parameters are effective properties that are related to the calibration system's scale and processes as well as the structure of the model, it is recommended that the type of data used for calibration is identical or similar to the type of the variable we are interested in predicting in a subsequent simulation run. The data should also

refer to a similar support scale to avoid upscaling or downscaling issues. If these two conditions are fulfilled, the model structure may not need to be modified substantially when changing from inverse to predictive modeling, making it more likely that model testing is successful.

For example, if we seek to determine the gas permeability of a fracture network on the continuum scale, it is reasonable to calibrate the model against gas pressure and gas tracer data collected in a large enough volume (e.g., in a ventilated tunnel) so that many fractures contribute to the observed signals (see Unger et al., 2004, for a discussion of this example). This is preferable over an approach that relies on a detailed geometric mapping of the fracture network combined with estimation of fracture permeabilities by calibrating data from water flow tests in a laboratory setting. Similarly, as discussed by Finsterle (2000), a model intended to predict dripping of water from an unsaturated fractured formation into a tunnel is best calibrated based on data that reflect this seepage process on the scale of a few meters. Moreover, the flow rates used in the testing should mimic (or at least approach) natural percolation conditions, which are the relevant fluxes for predicting seepage rates. Note that, even in this case, substantial extrapolation from the conditions encountered during calibration to those used in subsequent simulations is required. Nevertheless, following the principle of matching scale, process, and data type helps the modeler arrive at a defensible prediction model.

In apparent violation of this principle, one may take advantage of the coupled nature of many subsurface processes to estimate parameters based on data that only have an indirect effect on the target behavior. For example, while it is reasonable to estimate thermal properties using temperature data and two-phase properties using saturation or water potential data, temperature data may also be used to estimate two-phase parameters through the indirect effect of gas saturation on heat conductivity (Zhang et al., 2011). Other examples include the estimation of percolation flux by measuring temperature profiles, which includes a convective component superimposed on the conductive geothermal gradient (Bodvarsson et al., 2003). If a model accurately accounts for such coupled processes, the data base available for model calibration increases dramatically, as secondary effects (which are often considered noise in the data) become signals that can be exploited for parameter estimation.

A second principle addresses parameter correlations. As will be discussed in the next section, even a perfect match to observed data does not guarantee that the estimated parameters are meaningful. The estimates may be highly uncertain as a result of strong parameter correlations. While this issue may be considered a parameterization problem, it can be alleviated by choosing an appropriate test design and collecting data that contain independent, complementary information about the parameters of interest. For example, steady-state drawdown data do not contain enough information to estimate both the formation's conductivity and storativity; only the ratio of the two, i.e., the hydraulic diffusivity, can be estimated reliably. Performing a tracer test and adding tracer breakthrough curves to the calibration data set would provide independent information to help reduce the correlation between these two parameters. Note, however, that the tracer transport velocity depends on additional parameters (specifically porosity) that may introduce indirect correlations (e.g., among porosity, storativity, and conductivity). The ability of additional data sets to resolve parameter correlations can be examined prior to data collection as part of a test design that involves synthetic joint data inversions. Proposing alternative test designs requires a good understanding of the physics of the problem. An example is discussed by Finsterle and Persoff (1997), where the strong correlation between permeability and the Klinkenberg parameter can be broken by jointly inverting data from multiple gas pressure pulse decay experiments conducted on various pressure levels.

A third principle is concerned with the fact that any error in the conceptual model leads to an error in the estimated parameters. It is thus desirable that data are collected that also contain information about the model structure. This can be borehole data introduced as conditioning points or prior information. More recently, the joint inversion of hydrogeological and geophysical data has been proposed as a way to include high-resolution information about the structure of the subsurface into the model (Rubin and Hubbard, 2005). The basic idea is that geophysical data help resolve geometric features, while hydrogeological data (as well as thermal, geochemical, and geomechanical data) help identify the physical (as well as thermal, chemical, and mechanical) properties. It is essential that all these data types are inverted concurrently, which requires coupled forward models and potentially the estimation of additional parameters (for example, those of a petrophysical model, which relates the hydrological system state to geophysical attributes). Joint hydrogeophysical inversions performed by iTOUGH2 are described by Kowalsky et al. (2004, 2005, 2008, 2011) and Finsterle and Kowalsky (2008).

The observations to be used in an iTOUGH2 analysis need to be mapped to one or several output variables of the forward model, and they may be transformed (e.g., using a Box-Cox transformation; see Finsterle and Zhang (2011b) for details) before being assembled in a vector $\mathbf{z}$ of length m, where $\mathbf{z}^*$ holds the measured data and $\mathbf{z}(\mathbf{p})$ contains the corresponding model output, which is a function of the parameter vector $\mathbf{p}$. Note that the observation vector $\mathbf{z}$ may also contain prior information about the parameters, other regularization expressions, and penalty terms. The residual vector is defined as $\mathbf{r} = (\mathbf{z}^* - \mathbf{z}(\mathbf{p}))$.

It must be kept in mind that data collection is likely the most expensive part of site characterization. Careful sensitivity analyses and synthetic inversions can help design experiments and monitoring systems so they yield observation data that can be used for a successful calibration of the model and the determination of parameters with acceptably low estimation uncertainty.

iTOUGH2 SUPPORT OF CALIBRATION AND UNERTAINTY ANALYSIS

The iTOUGH2 simulation-optimization software (Finsterle, 1997a, 1997b, 1997c; http://esd.lbl.gov/iTOUGH2) supports the development, calibration, and testing of TOUGH models (Pruess et al., 1999; http://esd.lbl.gov/TOUGH). iTOUGH2 can also be used in connection with other forward models (Finsterle and Zhang, 2011a). iTOUGH2 has three main application modes: (1) formal local and global sensitivity analysis, (2) parameter estimation by automatic model calibration, and (3) linear uncertainty propagation analysis and Latin hypercube Monte Carlo simulations. As discussed above, essentially any TOUGH input parameter can be estimated or analyzed based on any type of observable data for which a corresponding TOUGH output variable is calculated.

Sensitivity Analysis

iTOUGH2 can be used to perform a local sensitivity analysis, i.e., it evaluates the partial derivative of select output variables with respect to select input parameters. Since the units, range of numerical values, and expected variability of both the parameters and observations can be very different within the same analysis, it is suggested to evaluate a normalized local sensitivity coefficient:

$$\Omega_{ij} = \frac{\partial z_i}{\partial p_j} \cdot \frac{\sigma_{p_j}}{\sigma_{z_i}} = J_{ij} \cdot \frac{\sigma_{p_j}}{\sigma_{z_i}} \qquad (1)$$
$$i = 1,\ldots,m;\ j = 1,\ldots,n$$

where σ_p is the parameter variation, reflecting the parameter's expected or acceptable uncertainty, and σ_z is the standard deviation of the observation, reflecting its measurement error or acceptable mean residual in an inversion. The sensitivity coefficients are summarized in a matrix $\mathbf{\Omega}$ of dimensions $m \times n$, where m is the number of observations and n is the number of parameters. The matrix holding the partial derivatives is termed the Jacobian matrix, $\mathbf{J} = (\partial \mathbf{z}/\partial \mathbf{p}) = -(\partial \mathbf{r}/\partial \mathbf{p})$. The partial derivatives are calculated numerically; the (n+1) or (2n+1) forward simulations needed to evaluate the sensitivity coefficients using forward or centered finite differences, respectively, can be performed in parallel (Finsterle, 1998). iTOUGH2 calculates composite sensitivity measures (e.g., the sum of the absolute values of the scaled sensitivity coefficients of each row and column of $\mathbf{\Omega}$, or separated for each data set and data type), which contain information about the overall influence of a parameter or the overall sensitivity of a data point, data set, or data type.

In addition, iTOUGH2 can be used to perform a global sensitivity analysis, which may be essential given the highly nonlinear character of most TOUGH models. The methods of Morris (1991) and Saltelli et al. (2006) are implemented. In the elementary effects method of Morris (1991), a point in the parameter space is selected, each parameter is perturbed, one at a time, and the corresponding impact on the output is evaluated. The procedure is repeated for multiple, particularly selected points in the parameter space. The mean of all elementary effects assesses the overall influence of the respective parameter on the output; the standard deviation indicates whether the effects are linear and additive or nonlinear, or whether interactions among the parameters are involved.

The variance-based method of Saltelli et al. (2006) uses sampling matrices to evaluate which part of the output variance can be explained by which input parameter, and how much the output variance can be reduced by fixing a parameter. This provides first-order sensitivity measures for each parameter.

The elementary effect method of Morris (1991) is an efficient approach that provides valuable information about a parameter's relative influence for nonlinear models. It requires a significantly smaller sample size than the variance-based method of Saltelli et al. (2006), which makes it very attractive for global sensitivity analyses for computationally expensive TOUGH models. It is highly recommended to perform a sensitivity analysis when designing laboratory and field experiments, and as a parameter screening tool prior to model calibration. However, one should keep in mind that high overall sensitivity is a necessary, but not sufficient, requirement for successful parameter estimation by inverse modeling.

Parameter Estimation

The main iTOUGH2 application mode is parameter estimation by automatic model calibration. An overall measure of misfit between the observed and calculated system response is minimized by automatically adjusting input parameters, and an extensive residual and uncertainty analysis is performed before the optimal parameter set is reported in the output file. The typical misfit criterion to be minimized is the standard weighted least squares objective function:

$$S = \mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r} \qquad (2)$$

where $\mathbf{C}_{zz}$ is the $m \times m$ covariance matrix of the observations. It contains the modeler's expectations about the size of the final residuals. If the forward model were a perfect representation of the true physical system, $\mathbf{C}_{zz}$ could be set identical to the covariance matrix of the measurement errors. However, since any model is by definition a simplified abstraction of the true system, modeling errors should be taken into account. They often lead to covariance matrices with relatively large off-diagonal terms (Lehikoinen et al., 2010). The inverse of $\mathbf{C}_{zz}$ serves as the weighting matrix. It accounts for data of dif-

ferent size and measurement units as well as different quality. It describes the random, non-identifiable component of the observation and is thus referred to as the stochastic model of the inversion. By contrast, the functional model captures the systematic, identifiable component of an observed system response; it is the responsibility of the physics-based forward model (the TOUGH model in our case) to capture this aspect. Inverse modeling attempts to extract information about the model parameters from the systematic part of the data, not from the random component. It is thus essential to identify and avoid overparameterization and overfitting.

Using the weighted least squares objective function leads to maximum likelihood estimates, provided that the residuals are normally distributed. While this assumption may be justified in many cases due to the central limit theorem, the distribution of residuals is often non-Gaussian and non-symmetric as a result of systematic modeling and measurement errors. In particular, they often exhibit tails that are more pronounced than those predicted by the normal distribution, which leads to biased estimates if the least-squares objective function is used. It is therefore essential to perform a careful analysis of the final residuals (see discussion below). Moreover, iTOUGH2 offers alternative objective functions, referred to as robust estimators; they are discussed by Finsterle and Najita (1998) and Finsterle and Zhang (2011b).

The topology of the objective function in the *n*-dimensional parameter space defines the inverse problem in its entirety, including whether it is ill- or well-posed, what the influence and correlation of the parameters are, and what the estimation uncertainty will be. The purpose of the optimization algorithm is to find the minimum of this objective function using a limited number of model evaluations. Examining the entire feasible parameter space is computationally prohibitive, especially if more than three parameters are estimated. Many minimization algorithms have been proposed in the literature and implemented in general software libraries or application-specific codes. Each of these algorithms makes certain assumptions about the topology of the objective function, such as whether the function is continuous or discontinuous, whether it is convex or exhibits multiple local minima, whether it is nearly quadratic or highly nonlinear, and whether no derivatives or first or second derivatives are available. Algorithms that are based on more restrictive assumptions are naturally more efficient than those that do not make any assumptions.

Several minimization algorithms have been implemented into iTOUGH2. The available derivative-free methods include (1) a simple grid search, where the objective function is evaluated on a regular grid or at user-specified sampling points, (2) the downhill simplex algorithm (Nelder and Mead, 1965; Finsterle, 2005), and several metaheuristic global minimization algorithms, namely (3) a differential evolutionary algorithm (Storn and Price, 1997), which mimics evolution with weighted differences between populations and a trial vector, (4) simulated annealing (Metropolis et al., 1953), which mimics the slow cooling of metals, (5) the harmony search algorithm (Ayvaz, 2007), which mimics musical improvisation, and (6) the multistart metric stochastic radial basis function approach (Regis and Shoemaker, 2007), which proposes new parameter combinations based on the distance from previous sets and the value of the objective function as predicted by an approximation of the response surface using radial basis functions. While these global search algorithms are potentially very powerful, they tend to require many solutions of the forward problem, as each method includes random steps (depending on the analogy, these steps are termed mutation and crossover, temperature fluctuation, pitch adjustment, etc.). Moreover, the algorithms require specification of problem-specific heuristic metaparameters that are difficult to determine *a priori*. Given that a typical TOUGH forward simulation is computationally very expensive, more efficient, derivative-based minimization algorithms need to be used for most practical applications at the expense of generality. The Levenberg-Marquardt modification of the Gauss-Newton algorithm has proven a flexible, robust, and efficient method to calibrate highly nonlinear TOUGH models. The Gauss-Newton algorithm is a second-order method in which the Hessian matrix is approximated by the Fisher information matrix, $\mathbf{J}^T\mathbf{C}_{zz}{}^{-1}\mathbf{J}$. This approximation is valid for linear models, weakly nonlinear models, or near the minimum; in these cases, the Gauss-Newton method can identify the minimum in a single iteration. For strongly nonlinear models, however, the Hessian is not necessarily a positive-definite matrix, and its approximation by the Fisher information matrix may not lead to an efficient or successful step. In the Levenberg-Marquardt method, the approximation to the Hessian is made positive definite by adding an $n \times n$ diagonal matrix $\lambda_k \mathbf{D}_k$ for a parameter update $\Delta\mathbf{p}_k$ at iteration k of:

$$\Delta\mathbf{p}_k = \left(\mathbf{J}_k^T\mathbf{C}_{zz}^{-1}\mathbf{J}_k + \lambda_k\mathbf{D}_k\right)^{-1}\mathbf{J}_k^T\mathbf{C}_{zz}^{-1}\mathbf{r}_k \qquad (3)$$

The scalar λ is the so-called damping or Levenberg parameter (Levenberg, 1944). If λ_k is zero, $\Delta\mathbf{p}_k$ is identical to the Gauss-Newton step; as $\lambda_k \to \infty$, the approximation of the Hessian becomes diagonally dominant, $\Delta\mathbf{p}_k$ becomes parallel to the steepest-descent direction, and the step length approaches zero. After each iteration, the Levenberg parameter is either increased or decreased following a scheme proposed by Marquardt (1963). For the initial iterations far away from the minimum, a relatively large value of λ_k is chosen (or automatically assigned by iTOUGH2 based on the initial value of the objective function), leading to small steps along the gradient of the objective function. After each successful step, i.e., if $S(\mathbf{p}_{k+1}) < S(\mathbf{p}_k)$, the Levenberg parameter λ_k is decreased by a factor of $1/\nu$, where $\nu > 1$ is the so-called Marquardt parameter. With decreasing λ_k, $\Delta\mathbf{p}_k$ approaches the Gauss-Newton step with its quadratic convergence rate. If an unsuccessful step is proposed, i.e., the objective function is increased, λ_k is increased by ν and a new trial set is tested. iTOUGH2 provides three options for the Tikhonov matrix $\mathbf{D}$ in equation 3 and allows for a dynamic truncation of the parameter space. In this scheme, only parameters that have sufficient support from the observed data are updated during a particular iteration.

Due to the nonlinearity of the problem, parameters may be dynamically included or excluded from the solution space as the optimization proceeds (for details, see Finsterle and Kowalsky, 2011). Optimization may also proceed along a reduced set of so-called superparameters (which are aligned with the eigenvectors of the Fisher information matrix), an approach similar to that proposed by Tonkin and Doherty (2005).

As mentioned above, the choice of a suitable minimization algorithm should be guided by the properties of the objective function and the available computer resources. It is obvious that even the best minimization algorithm does not result in meaningful parameter estimates if the inverse problem is ill-posed. It is therefore highly recommended to invest in a careful design of the experiment, field test, or monitoring network, and to employ proper parameterization to make sure that the data available for model calibration allow the formulation of a well-posed inverse problem and the use of an efficient, second-order minimization algorithm.

Once the minimum of the objective function has been identified, a residual and uncertainty analysis must be performed. The parameter set at the minimum may be meaningless if the model is unlikely to represent the real system, or if the estimation uncertainty is large. The fit to the data can be considered acceptable if the overall goodness-of-fit criterion, expressed by the estimated error variance, i.e.:

$$s_0^2 = \frac{\mathbf{r}^T \mathbf{C}_{zz}^{-1} \mathbf{r}}{m-n} \tag{4}$$

is consistent with the modeler's expectation, which is expressed *a priori* through the observation covariance matrix $\mathbf{C}_{zz}$. This consistency can be statistically tested. However, even if the overall fit is deemed acceptable, the residuals may be unacceptable for certain portions of the data (e.g., at early times or at the peak values of a time series). A detailed residual analysis can shed light on an inconsistency between the data and the model, which is often apparent by the fact that the residuals (1) are not symmetric, (2) show a systematic trend rather than being random, (3) exhibit outliers, or (4) are significantly larger or smaller than expected. In many cases, a careful analysis of the residuals points toward aspects of the model that need to be refined. The statistical tools implemented in iTOUGH2 that support the residual analysis are discussed in detail by Finsterle and Zhang (2011b).

Even if the model is capable of accurately reproducing the observed system response at the calibration points, the parameter set may still be of little or no practical use because of large estimation uncertainty. Under the assumption that the forward model is linear and the residuals are Gaussian, the covariance matrix of the estimated parameters is given by:

$$\mathbf{C}_{pp} = s_o^2 (\mathbf{J}^T \mathbf{C}_{zz}^{-1} \mathbf{J})^{-1} \tag{5}$$

The variances of the estimated parameters may be large if (1) the model does not fit the data well enough, i.e., the estimated error variance or the elements of $\mathbf{C}_{zz}$ are too large, (2) the observations are not sufficiently sensitive to the parameters of interest, i.e., the values in a column of the Jacobian matrix $\mathbf{J}$ are too small, or (3) the parameters exhibit strong statistical correlations, i.e., the columns of $\mathbf{J}$ are (almost) linearly dependent, which leads to large off-diagonal elements in $\mathbf{C}_{pp}$. Strong parameter correlation is the main culprit of large estimation uncertainty and is usually an indication of overparameterization, a poor choice of the parameter set, or a lack of calibration data with complementary information content. This is demonstrated by Finsterle and Persoff (1997). While the covariance matrix $\mathbf{C}_{pp}$ of equation 5 is only an approximation of the actual uncertainty region if the model is nonlinear, it provides very useful information about the estimation uncertainty and parameter correlations (Finsterle and Pruess, 1995), which in turn can be used to improve the design of characterization experiments.

Finally, iTOUGH2 evaluates model identification criteria, which can be used as performance measures among competing alternative conceptual models with different parameterizations and different calibration data (Carrera and Neuman, 1986; Poeter and Anderson, 2005; Ye et al., 2008).

Uncertainty Analysis

Because the parameters estimated by inverse modeling are strictly valid only for the scale, processes, and model structure used by the calibration model, it is necessary to demonstrate that these parameters are also meaningful in a different context, i.e., if used for predictions under changed conditions. Uncertainty analysis is part of a confidence building process, which is sometimes referred to as model validation (see Oreskes et al., 1994, for a critical discussion of this term). It generally consists of comparing model predictions with measured data that were not used for calibration in an attempt to demonstrate that the model is capable of making reasonable predictions if applied within the framework of its intended use. For this comparison with measured data (or for any other simulation with or without corresponding measurements), it is necessary or desirable to calculate the uncertainty of the model prediction. The simplest way to calculate prediction uncertainty is to perform a linear or first-order second-moment error analysis:

$$\mathbf{C}_{\hat{z}\hat{z}} = \mathbf{J}\mathbf{C}_{pp}\mathbf{J}^T \tag{6}$$

where $\mathbf{C}_{\hat{z}\hat{z}}$ is the covariance matrix of the model predictions, $\mathbf{C}_{pp}$ is the parameter covariance matrix, and $\mathbf{J}$ is the Jacobian

matrix that relates a change in the input parameters to each output variable of interest. Note that additional parameters may be included in this uncertainty propagation analysis, i.e., the dimensions of $\mathbf{C}_{pp}$ may be larger than n.

While this uncertainty propagation analysis is efficient, it is based on an assumption of linearity of the forward model and normality of the residuals and may thus assign (hopefully small) probabilities to a predicted system behavior that is physically unfeasible. Monte Carlo simulations, in which the input distributions of the uncertain parameters are sampled many times and the corresponding output is described in statistical terms, is a very flexible way to account for any type of input distribution as well as nonlinearities. However, the relatively large number of model evaluations needed to obtain converged output statistics make Monte Carlo analyses computationally expensive, even though all of these evaluations are independent and therefore can be run in parallel. Advanced sampling designs (e.g., Latin hypercube sampling) can be used to make sure the samples better span the entire parameter space even with a limited number of sampling points.

iTOUGH2 supports linear uncertainty propagation analysis as well as Monte Carlo methods using random or Latin hypercube sampling. Correlations among the parameters can be accounted for using the approaches described by Zhang and Pinder (2003) and Kitterød and Finsterle (2004).

It is important to realize that the methods described here only consider parametric (epistemic) uncertainty; aleatoric uncertainties or uncertainties due to alternative model conceptualization are usually not accounted for, even though they often have a substantial effect on model predictions. Assigning weights and sampling of alternative conceptual models is not a trivial issue (Neuman, 2003). One approach is to parameterize conceptual model elements or to generate geostatistical realizations of heterogeneous property fields as part of each Monte Carlo evaluation to account for spatial variability; an example of this methodology is described by Finsterle (2000).

Prediction uncertainty and data uncertainty need to be combined to yield a measure that can be compared to an acceptance criterion used for model testing. A detailed discussion of model validation, however, is beyond the scope of this article. The journal *Advances in Water Resources* dedicated two issues to the topic of validation of groundwater models (AWR, 1992); the critical article by Oreskes et al. (1994) should also be considered.

CASE STUDIES

References to calibration studies using the iTOUGH2 simulation-optimization code have been given throughout the preceding sections; additional references can be found at http://esd.lbl.gov/iTOUGH2/Bibliography/bibliography.html. iTOUGH2 has been applied to the inversion of synthetic data for testing optimization algorithms and analysis tools, and to actual data from many small-scale laboratory experiments, plot-scale field tests, and large-scale site characterization efforts. Many of the published case studies emphasize the system behavior and the insights that can be gained by using a sophisticated nonisothermal multiphase flow and transport simulator capable of modeling coupled processes. In other words, the studies focus on the first-order effects as revealed by the forward model. In addition, calibration issues, estimation uncertainty, and the appropriateness of using a specific approach to solve a given inverse problem are discussed. Only two examples are presented here.

CASE STUDY 1: JOINT HYDROGEOPHYSICAL INVERSION

The first case study demonstrates the calibration of a three-dimensional, highly heterogeneous, unsaturated flow model that simulates infiltration of water into the vadose zone at the U.S. Department of Energy's Hanford site in Washington. The model is calibrated using a joint hydrogeophysical inversion, where time-lapse ground-penetrating radar (GPR) travel times and neutron probe (NP) data are concurrently matched to estimate field-scale soil hydraulic properties along with petrophysical parameters, which relate soil porosity and water saturation to the effective dielectric constant. The study is described in detail by Kowalsky et al. (2005).

At the Hanford site, highly radioactive waste and other toxic fluids have leaked into the vadose zone (e.g., Sisson and Lu, 1984) and have necessitated the development of methods for monitoring and ultimately controlling the spread of contamination. At the Hanford 200 East Area field site, an infiltration experiment was performed in May 2000 consisting of five injections of 4000 L of water each over a period of one month. During the experiment, cross-borehole GPR and NP measurements were made. Figure 1a shows the distribution of water content derived from the dense NP data cube. The locations of the wells in which cross-borehole GPR measurement data were collected are indicated in figure 1b, as are the straight-ray paths for the GPR measurements.

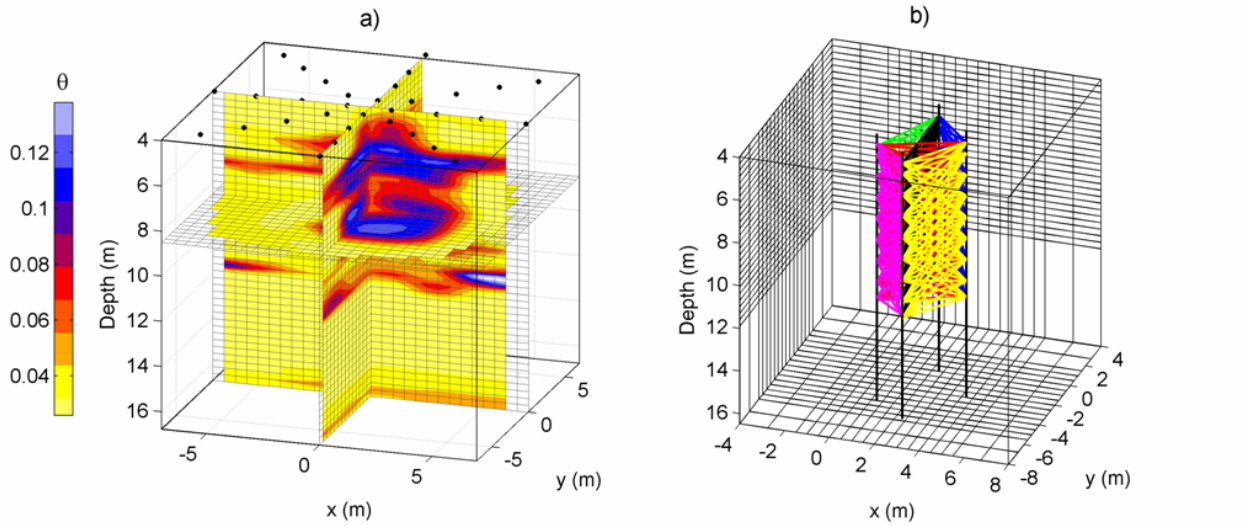

Figure 1. Available data sets collected during injection experiment at the Hanford site: (a) example of a densely sampled NP data set (interpolated from NP wells, the tops of which are indicated by black dots), and (b) straight-ray paths formed by the four GPR wells used for inversion.

To simulate the infiltration experiment and to estimate the distribution of hydrological properties, a three-dimensional TOUGH2 model with dimensions 12 m × 12 m × 14 m is developed. A small flux of water is applied at the top of the model to represent average infiltration conditions. Since the model domain is in the unsaturated zone far above the water table, a free drainage boundary is implemented at the bottom of the model domain. No-flow boundaries are set at the four vertical sides. For the injection source, a time-dependent mass flow rate is specified. Based on geostatistical analyses of permeability measurements, an anisotropic spherical semivariogram is chosen to model the log-permeability. In addition, GPR travel times are calculated using the straight-ray method. A petrophysical model is used to evaluate the electromagnetic velocity as a function of the TOUGH2-calculated water content, porosity, and the dielectric constants of the liquid, gas, and solid phases.

The parameter set to be estimated by automatic model calibration consists of the petrophysical parameter κ_s, which is the dielectric constant for the solid components, the porosity ϕ, the reference log-permeability in horizontal and vertical directions, the log-permeability modifier values at 16 pilot point locations, and a factor by which the reported flow rates at the injection point are multiplied. Using two GPR travel time surveys and a small subset of the available NP data, multiple inversions with different realizations of the initial log-permeability field are performed. Figure 2 shows a vertical profile of predicted and measured water content at a location near the injection well, along with the prediction uncertainty bounds. Figure 3 shows one example of the log-permeability distribution estimated by concurrently fitting GPR arrival times and NP data in a joint inversion framework. The forward model consists of the hydrological and geophysical models linked together through the petrophysical relationship.

This case study demonstrates that the complementary information contained in geophysical and hydrological data can be successfully extracted in a joint inversion approach. Moreover, since no tomogram (an image of electromagnetic velocity derived from a traditional crosshole tomography procedure) needs to be generated, the amount of GPR data required for analysis is relatively low, and difficulties inherent in tomography methods are alleviated. Finally, the approach provides a means to capture the properties and system state of heterogeneous soil, both of which are crucial for assessing and predicting subsurface flow and contaminant transport.

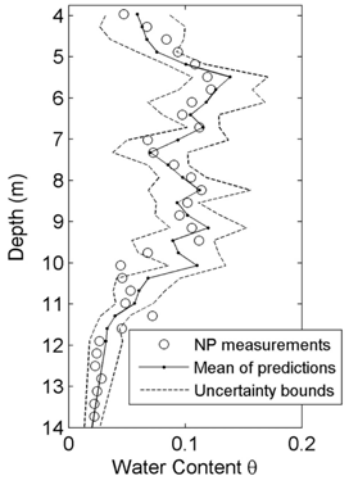

Figure 2. Comparison of water content profile near the injection point 15 days after infiltration started, obtained from the dense NP data set (circles) with the mean of the predicted water content profiles (solid line with dots) obtained through inversion of both the GPR data set and the limited NP data set. The dashed lines indicate the estimation uncertainty (±2 standard deviations).

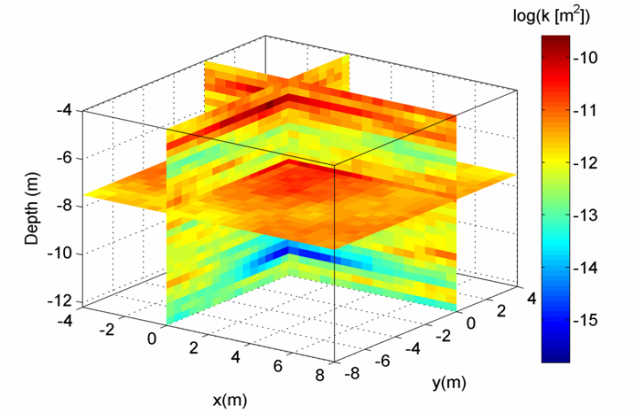

Figure 3. Log-permeability distribution estimated by the joint inversion of geophysical and hydrological data. Permeability modifiers are estimated at 16 pilot points, and geostatistical simulations are performed to obtain the spatially correlated property field.

CASE STUDY 2: TWO-CRITERIA OPTIMIZATION PROBLEM FOR AQUIFER REMEDIATION

The second example demonstrates the use of iTOUGH2 for the evaluation of a Pareto frontier. The Pareto frontier can be considered the set of solutions to a multicriteria optimization problem, where the relative weights of the criteria are varied to examine the tradeoffs between competing objectives. Here, we determine the Pareto frontier by running multiple iTOUGH2 inversions, where the relative weights are adjusted in predefined increments. The weights are those assigned to two observation types, each representing a different objective. For each weight combination, an iTOUGH2 inversion is performed, and the mean residual of each observation type is extracted and used to create the Pareto frontier plot. In this example, an outer iTOUGH2 run controls other iTOUGH2 optimization runs through the PEST protocol (Finsterle and Zhang, 2011a).

The optimization problem considered is a synthetic remediation design problem in which the tradeoff between two objectives is examined. These competing objectives are (1) maximization of contaminant removal within a specified cleanup time of five years, and (2) minimization of cleanup costs, simplified here as the total amount of water pumped from six wells during a pump-and-treat operation. The individual minimization problem of determining optimal pumping rates (assuming that the relative costs of pumping and residual contamination are known) is summarized here and described in detail by Finsterle (2005). In a simplified scenario, an aquifer contaminated by trichloroethylene (TCE) will be remediated by drilling six wells approximately aligned with the plume axis. It is then assumed that the costs for site remediation are directly proportional to the total amount of contaminated water being extracted from the subsurface. In other words, the

cost function to be minimized consists of two components: (1) the sum of the estimated pumping rates at the six wells, and (2) the amount of TCE left in the aquifer after the five-year cleanup period. (The relative weighting of these two components of the cost function is discussed below.) The cost function can be successfully minimized using any of the algorithms available in iTOUGH2 (Finsterle, 2005); the downhill simplex algorithm (Nelder and Mead, 1965) was used for this study. The forward problem is solved using the three-phase simulator T2VOC (Falta et al., 1995).

This optimization problem is then solved repeatedly for different weights of the two competing objectives. By giving higher weight to the remediation goal, pumping rates are expected to go up; conversely, if emphasis is placed on reducing pumping costs, the pumping rates will generally go down at the expense of increased residual contamination. The tradeoff between these two objectives is evaluated at 40 discrete points with relative weights (w_p and w_c) for the pumping cost and remediation objectives, respectively, under the constraint that $w_p + w_c = 1$. Only one of the weights (that of the pumping rate criterion, w_p) is varied from zero to one. The weight given to the residual contamination criterion is tied to the first parameter. For each weight combination, the optimal distribution of pumping rates in the six wells is determined by an iTOUGH2 optimization that minimizes both the (weighted) total amount of water pumped and the (weighted) residual contaminant mass. This six-parameter minimization problem is solved 40 times. The total pumping rate and the residual contaminant mass after each optimization is extracted from the residual analysis section of the iTOUGH2 output file. Plotting the two objectives against each other provides the Pareto frontier.

The resulting Pareto frontier is shown in figure 4, demonstrating that there is a relatively well-defined optimal solution (i.e., the region of the Pareto frontier near the origin), where both criteria can be met without too much tradeoff. This case study demonstrates that the minimization algorithms implemented in iTOUGH2 for model calibration can also be used to solve a broader spectrum of optimization problems. In this example, it determined optimum pumping rates that removed contaminants from an aquifer at minimum cost. Moreover, the flexibility provided by the PEST protocol allowed nesting of iTOUGH2 analysis tools so that a Pareto frontier could be calculated by invoking many optimization runs with different weighting of competing criteria.

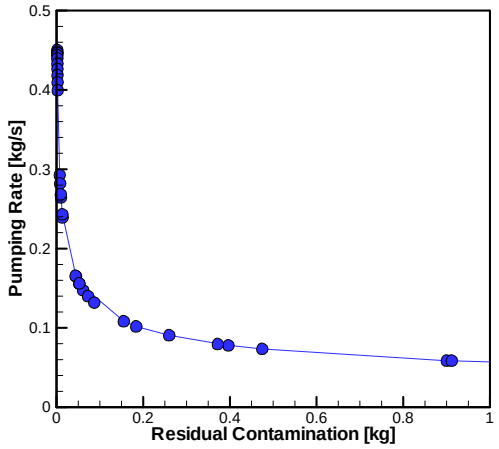

Figure 4. Pareto frontier of two-criteria optimization problem for efficient cleanup of contaminated groundwater aquifer. Multiple minimizations problems are solved using iTOUGH2 to determine pumping rates in six remediation wells.

DISCUSSION

The simulation of coupled processes in complex subsurface systems requires the estimation of a relatively large number of diverse parameters that represent effective formation properties and lumped processes that cannot be captured individually on the continuum scale. Automatic model calibration using relevant, complementary data of high sensitivity and good quality is a defensible approach to estimate such effective parameters. Because the objectives of a modeling study vary substantially among applications, it is not possible to provide a definite list of parameters that need to be estimated, nor to describe an ideal data set that should be used for calibration. We consider it essential that influential parameters and data needs for estimation be determined individually for each study, using sensitivity analyses, synthetic inversions, and detailed uncertainty analyses. These analyses can be done prior to data collection. Once testing or site characterization data are available for calibration, developing a robust forward model of appropriate complexity is a key step, as any error in the conceptual model leads to errors in the estimated parameters that are usually much larger than uncertainties introduced by random measurement noise. The forward model has to be capable of calculating output variables that are conceptually consistent with the observations. If the model is, in principle, able to reproduce the observed data by adjusting the parameters that were selected for analysis, the model can be calibrated. This requires the definition of the calibration points in

space and time and their relative weights, the selection of an objective function, and a suitable minimization algorithm. Inversion is not complete without a detailed residual and uncertainty analysis.

iTOUGH2 supports many of the tasks that are needed for the successful development, calibration, and testing of complex subsurface flow and transport models. Local and global sensitivity analyses can be used to evaluate data worth and for parameter screening. Local and global minimization algorithms are available for finding the minimum of the objective function. A detailed residual and uncertainty analysis is performed to examine whether the model is a reasonable representation of the observed system, and whether the estimation uncertainty is acceptable for subsequent use in a prediction model. Finally, linear and Monte Carlo-style uncertainty propagation analyses provide insights into the reliability of the model predictions and can be used as a basis for evaluating acceptance criteria in a confidence-building exercise.

The combination of the physics-based forward model TOUGH2 and the calibration and analysis tools of iTOUGH2 provide a comprehensive simulation-optimization package that has been successfully applied to a large variety of complex subsurface flow and transport problems. Many challenges and limitations remain; they are addressed in the next section.

FUTURE DEVELOPMENTS

Model calibration and predictive modeling are theoretically difficult and challenging in practice. In general, fundamental issues related to parameterization, uniqueness, stability, and uncertainty analysis need to be resolved. Efficient and robust minimization algorithms need to be developed that can handle highly parameterized, nonlinear inverse problems. Software packages need to be written that are user friendly without hiding the complexity inherent in these analyses. Finally, education and training in forward and inverse modeling is an essential prerequisite for the successful and defensible application of automatic model calibration and uncertainty quantification techniques in hydrogeology and environmental applications.

More specifically, we consider research efforts in the following areas necessary, as they provide promising avenues to improve the value that can be obtained from characterization data for the calibration of environmental models:

- Novel ways of parameterizing subsurface heterogeneity are needed. These methods must be flexible enough to capture the large variety of subsurface structures using a limited number of parameters. Moreover, the resulting structures must yield realistic features that are consistent with the general geologic information available at a site as well as with more quantitative data, such as well logs and geophysical data.
- Novel data collection approaches are needed that provide a large amount of relevant data of sufficient spatial resolution at acceptable cost. Geophysical surveys (from boreholes, the surface, or by remote sensing) as well as advances in sensor technology and data acquisition may provide the calibration data needed to solve otherwise ill-posed inverse problems.
- Novel algorithms need to be developed that allow the integration of large data sets of various types for the estimation of diverse parameters.
- Forward models need to be developed that are capable of simulating the observed data in a fully coupled manner.
- Model reduction methods need to be developed to significantly increase the efficiency of function evaluations during an inversion.
- Parameter screening and regularization methods need to be developed to stabilize the inversion without introducing an estimation bias.
- Computer codes need to be developed that support all aspects of model development. These codes should be made available to the scientific and engineering community; access to high-performance computing platforms may be offered.
- Best practices for experimental design, model development, calibration, and testing need to be formulated.

All these developments require strong interdisciplinary cooperation, with input from hydrogeologists, geochemists, geophysicists, computer scientists, applied mathematicians, statisticians, and other specialists, and need to be tailored to the needs of practitioners. Model calibration helps bridge the gap between fundamental process understanding and site-specific predictive modeling. Overcoming its challenges is thus essential to solve future engineering, environmental, and energy recovery problems.

ACKNOWLEDGEMENTS

We greatly appreciate the constructive and detailed comments on the manuscript by three anonymous reviewers. This work was supported, in part, by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Resources under Award No. DE-AC02-05CH11231.

REFERENCES

AWR. 1992. Special issue: Validation of geo-hydrological models: Parts 1 and 2. M. A. Celia, W. G. Gray, S. M. Hassanizadeh, and J. Carrera, eds. *Advances in Water Resources* 15(1 and 3).

Ayvaz, M. T. 2007. Simultaneous determination of aquifer parameters and zone structures with fuzzy c-means clustering and meta-heuristic harmony search algorithm. *Advances in Water Resources* 30(11): 2326-2338.

Bodvarsson, G. S., E. Kwicklis, C. Shan, and Y. S. Wu. 2003. Estimation of percolation flux from borehole temperature data at Yucca Mountain, Nevada. *J. Contam. Hydrol.* 62-63: 3-22.

Carrera, J., and S. P. Neuman. 1986. Estimation of aquifer parameters under transient and steady-state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resources Res.* 22(2): 199-210.

Doughty, C. 1999. Investigation of conceptual and numerical approaches for evaluating moisture, gas, chemical, and heat transport in fractured unsaturated rocks. *J. Contam. Hydrol.* 38(1-3): 69-106.

Edwards, A. L. 1972. TRUMP: A computer program for transient and steady-state temperature distributions in multidimensional systems. Springfield, Va.: National Bureau of Standards, National Technical Information Service.

Engelhardt, I., and S. Finsterle. 2003. Thermal-hydraulic experiments with bentonite/crushed rock mixtures and estimation of effective parameters by inverse modeling. *Applied Clay Sci.* 23(1-4): 111-120.

Falta, R. W., K. Pruess, and S. Finsterle. 1995. T2VOC user's guide. Report LBL-36400. Berkeley, Cal.: Lawrence Berkeley Laboratory.

Finsterle, S. 1992. iTOUGH2 user's guide, Version 1.0. Nagra internal report. Wettingen, Switzerland: National Cooperative for the Disposal of Radioactive Waste (Nagra).

Finsterle, S. 1997a. iTOUGH2 user's guide. Report LBNL-40040. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Finsterle, S. 1997b. iTOUGH2 command reference. Report LBNL-40041. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Finsterle, S. 1997c. iTOUGH2 sample problems. Report LBNL-40042. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Finsterle, S. 1998. Parallelization of iTOUGH2 using PVM. Report LBNL-42261. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Finsterle, S. 2000. Using the continuum approach to model unsaturated flow in fractured rock. *Water Resources Res.* 36(8): 2055-2066.

Finsterle, S. 2004. Multiphase inversion modeling: Review and iTOUGH2 applications. *Vadose Zone J.* 3(3): 747-762.

Finsterle, S. 2005. Demonstration of optimization techniques for groundwater plume remediation using iTOUGH2. *Environ. Modelling and Software* 21(5): 665-680, doi: 10.1016/j.envsoft.2004.11.012.

Finsterle, S., and B. Faybishenko. 1999. Inverse modeling of a radial multistep outflow experiment for determining unsaturated hydraulic properties. *Advances in Water Resources* 22(5): 431-444.

Finsterle, S., and M. B. Kowalsky. 2008. Joint hydrological-geophysical inversion for soil structure identification. *Vadose Zone J.* 7(1): 287-293, doi: 10.2136/vzj2006.0078.

Finsterle, S., and M. B. Kowalsky. 2011. A truncated Levenberg-Marquardt algorithm for the calibration of highly parameterized models. *Computers and Geosci.* 37(6): 731-738, doi: 10.1016/j.cageo.2010.11.005.

Finsterle, S., and J. Najita. 1998. Robust estimation of hydrogeologic model parameters. *Water Resources Res.* 34(11): 2939-2947.

Finsterle, S., and P. Persoff. 1997. Determining permeability of tight rock samples using inverse modeling. *Water Resources Res.* 33(8): 1803-1811.

Finsterle, S., and K. Pruess. 1995. Solving the estimation-identification problem in two-phase flow modeling. *Water Resources Res.* 31(4): 913-924.

Finsterle, S., and Y. Zhang. 2011a. Solving iTOUGH2 simulation and optimization problems using the PEST protocol. *Environ. Modelling and Software* 26(7): 959-968, doi: 10.1016/j.envsoft.2011.02.008.

Finsterle, S., and Y. Zhang. 2011b. Error handling strategies in multiphase inverse modeling. *Computers and Geosci.* 37(6): 724-730, doi: 10.1016/j.cageo.2010.11.009.

Finsterle, S., C. Doughty, M. B. Kowalsky, G. J. Moridis, L. Pan, T. Xu, Y. Zhang, and K. Pruess. 2008. Advanced vadose zone simulations using TOUGH. *Vadose Zone J.* 7(2): 601-609, doi: 10.2136/vzj2007.0059.

Freifeld, B. M., S. Finsterle, T. C. Onstott, T. Toole, and L. M. Pratt. 2008. Ground surface temperature reconstructions: Using *in situ* estimates for thermal conductivity acquired with a fiber-optic distributed thermal perturbation sensor. *Geophysical Res. Letters* 35: L14309, doi: 10.1029/2008GL034762.

Gallagher, P. M., and S. Finsterle. 2004. Physical and numerical model of colloidal silica injection for passive site stabilization. *Vadose Zone J.* 3(3): 917-925.

Ghezzehei, T. A., R. C. Trautz, S. Finsterle, P. J. Cook, and C. F. Ahlers. 2004. Modeling coupled evaporation and seepage in ventilated tunnels. *Vadose Zone J.* 3(3): 806-818.

Itasca. 1997. FLAC-3D manual: Fast Lagrangian analysis of continua in 3 dimensions: Version 2.0. Minneapolis, Minn.: Itasca Consulting Group, Inc.

Jung, Y., P. Imhoff, and S. Finsterle. 2011. Estimation of landfill gas generation rate and gas permeability of refuse using inverse modeling.

Transport in Porous Media 90(1): 41-58, doi: 10.1007/s11242-010-9659-8.

Kiryukhin, A. V., N. P. Asaulova, and S. Finsterle. 2008. Inverse modeling and forecasting for the exploitation of the Pauzhetsky geothermal field, Kamchatka, Russia. *Geothermics* 37(5): 540-562, doi: 10.1016/j.geothermics.2008.04.003.

Kitterød, N.-O., and S. Finsterle. 2004. Simulating unsaturated flow fields based on saturation measurements. *J. Hydraulic Res.* 42(supp. 1): 121-129.

Kowalsky, M. B., S. Finsterle, and Y. Rubin. 2004. Estimating flow parameter distributions using ground-penetrating radar and hydrological measurements during transient flow in the vadose zone. *Advances in Water Resources* 27(6): 583-599.

Kowalsky, M. B., S. Finsterle, J. Peterson, S. Hubbard, Y. Rubin, E. Majer, A. Ward, and G. Gee. 2005. Estimation of field-scale soil hydraulic parameters and dielectric parameters through joint inversion of GPR and hydrological data. *Water Resources Res.* 41: W11425, doi: 10.1029/2005WR004237.

Kowalsky, M. B., J. Birkholzer, J. Peterson, S. Finsterle, S. Mukhopadhyay, and Y. Tsang. 2008. Sensitivity analysis for joint inversion of ground-penetrating radar and thermal-hydrological data from a large-scale underground heater test. *Nuclear Tech.* 164(2): 169-179.

Kowalsky, M. B., S. Nakagawa, and G. J. Moridis. 2010. Feasibility of monitoring gas-hydrate production with time-lapse vertical seismic profiling. *Soc. Petroleum Eng. J.* 15(3): 634-645.

Kowalsky, M. B., E. Gasperikova, S. Finsterle, D. Watson, and S. S. Hubbard. 2011. Coupled modeling of hydrogeochemical and electrical resistivity data for exploring the impact of recharge on subsurface contamination. *Water Resources Res.* 47: W02509, doi: 10.1029/2009WR008947.

Kowalsky, M. B., S. Finsterle, K. H. Williams, C. Murray, M. Commer, D. Newcomer, A. Englert, C. I. Steefel, and S. S. Hubbard (in press). On parameterization of the inverse problem for estimating aquifer properties using tracer data.

Lehikoinen, A., J. M. J. Huttunen, S. Finsterle, M. B. Kowalsky, and J. P. Kaipio. 2010. Dynamic inversion for hydrological process monitoring with electrical resistance tomography under model uncertainties. *Water Resources Res.* 46: W04513, doi: 10.1029/2009WR008470.

Levenberg, K. 1944. A method for the solution of certain nonlinear problems in least squares. *Qtly. J. Applied Math.* 2(2): 164-168.

Marquardt, D. W. 1963. An algorithm for least squares estimation of nonlinear parameters. *SIAM J. Applied Math.* 11(2): 431-441.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Physics* 21(6): 1087-1092.

Moore, C., and J. Doherty. 2006. The cost of uniqueness in groundwater model calibration. *Advances in Water Resources* 29(4): 605-623.

Moridis, G. J. 2003. Numerical studies of gas production from methane hydrates. *Soc. Petroleum Eng. J.* 32(8): 359-370.

Moridis, G., and K. Pruess. 1998. T2SOLV: An enhanced package of solvers for the TOUGH2 family of reservoir simulation codes. *Geothermics* 27(4): 415-444.

Moridis, G. J., M. B. Kowalsky, and K. Pruess. 2008. TOUGH+HYDRATE v1.0 user's manual: A code for the simulation of system behavior in hydrate-bearing geologic media. Report LBNL-149E. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Morris, M. D. 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics* 33(2): 161-174.

Narasimhan, N. T., and P. A. Witherspoon. 1976. An integrated finite difference method for analyzing fluid flow in porous media. *Water Resources Res.* 12(1): 57-64.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. *Computer J.* 7(4): 307-313.

Oldenburg, C. M., and K. Pruess. 2000. Simulation of propagating fronts in geothermal reservoirs with the implicit Leonard total variation diminishing scheme. *Geothermics* 29(1): 1-25.

Neuman, S. P. 2003. Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models. *Stochastic Environ. Res. and Risk Assess.* 17(5): 291-305, doi: 10.1007/s00477-003-0151-7.

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the Earth sciences. *Science* 264(5147): 641-646.

Poeter, E. P., and D. A. Anderson. 2005. Multimodal ranking of interference in ground water modeling. *Ground Water* 43(4): 597-605.

Pruess, K. 1983. Development of the general-purpose simulator MULKOM. Report LBL-15500. Berkeley, Cal.: Lawrence Berkeley Laboratory.

Pruess, K. 1987. TOUGH user's guide. Report LBL-20700. Berkeley, Cal.: Lawrence Berkeley Laboratory.

Pruess, K. 1991. TOUGH2: A general-purpose numerical simulator for multiphase fluid and heat flow. Report LBL-29400. Berkeley, Cal.: Lawrence Berkeley Laboratory.

Pruess, K. 2004. The TOUGH codes: A family of simulation tools for multiphase flow and transport processes in permeable media. *Vadose Zone J.* 3(3): 738-746.

Pruess, K. 2005. ECO2N: A TOUGH2 fluid property module for mixtures of water, NaCl, and CO_2. Report LBNL-57952. Berkeley, Cal.:

Lawrence Berkeley National Laboratory.

Pruess, K. 2011. ECO2M: A TOUGH2 fluid property module for mixtures of water, NaCl, and CO_2, including super- and sub-critical conditions, and phase change between liquid and gaseous CO_2. Report LBNL-4590E. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Pruess, K., and A. Battistelli. 2002. TMVOC: A numerical simulator for three-phase non-isothermal flows of multicomponent hydrocarbon mixtures in saturated-unsaturated heterogeneous media. Report LBNL-49375. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Pruess, K., C. Oldenburg, and G. Moridis. 1999. TOUGH2 user's guide, Version 2.0. Report LBNL-43134. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Regis, R. G., and C. A. Shoemaker. 2007. A stochastic radial basis function method for the global optimization of expensive functions. *INFORMS J. Computing* 19(4): 497-509.

Rubin, Y., and S. Hubbard, eds. 2005. *Hydrogeophysics*. Water and Science Technology Library, Vol. 50. Amsterdam, The Netherlands: Springer.

Rutqvist, J., Y.-S. Wu, C.-F. Tsang, and G. Bodvarsson. 2002. A modeling approach for analysis of coupled multiphase fluid flow, heat transfer, and deformation in fractured porous rock. *Intl. J. Rock Mech. and Mining Sci.* 39(4): 429-442.

Saltelli A., M. Ratto, S. Tarantola, and F. Campolongo. 2006. Sensitivity analysis practices: Strategies for model-based inference. *Reliability Eng. and System Safety* 91(10-11): 1109-1125.

Sisson, J. B., and A. H. Lu. 1984. *Field Calibration of Computer Models for Applications to Buried Liquid Discharges: A Status Report.* Tech. Rep. RHO-ST-46-P. Richland, Wash.: Rockwell Hanford Operations.

Storn, R., and K. Price. 1997. Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optimization* 11(4): 341-359.

Tonkin, M., and J. Doherty. 2005. A hybrid regularized inversion methodology for highly parameterized models. *Water Resources Res.* 41(10): W10412, doi: 10.1029/2005WR003995.

Unger, A., S. Finsterle, and G. S. Bodvarsson. 2004. Transport of radon gas into a tunnel at Yucca Mountain: Estimating large-scale fractured tuff hydraulic properties and implications for the ventilation system. *J. Contam. Hydrol.* 70(3-4): 152-171.

Xu, T., and K. Pruess. 2001. Modeling multiphase non-isothermal fluid flow and reactive geochemical transport in variably saturate fractured rocks: 1. Methodology. *American J. Sci.* 301(1): 16-33.

Xu, T., E. Sonnenthal, N. Spycher, and K. Pruess. 2004. TOUGHREACT user's guide: A simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media. Report LBNL-55460. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Ye, M., P. D. Meyer, and S. P. Neuman. 2008. On model selection criteria in multimodal analysis. *Water Resources Res.* 44: W03428. doi: 10.1029/2008WR006803.

Zhang, K., Y.-S. Wu, and K. Pruess. 2008. User's guide for TOUGH2-MP: A massively parallel version of the TOUGH2 code. Report LBNL-315E. Berkeley, Cal.: Lawrence Berkeley National Laboratory.

Zhang, Y., and G. Pinder. 2003. Latin hypercube lattice sample selection strategy for correlated random hydraulic conductivity fields. *Water Resources Res.* 39(8): 1226, doi: 10.1029/2002WR001822.

Zhang, Y., B. Freifeld, S. Finsterle, M. Leahy, J. Ennis-King, L. Paterson, and T. Dance. 2011. Single-well experimental design for studying residual trapping of supercritical carbon dioxide. *Intl. J. Greenhouse Gas Control* 5(1): 88-98, doi: 10.1016/j.ijggc.2010.06.011.