ground water

A Methodology for Validating Numerical Ground Water Models

by Ahmed E. Hassan[1]

Abstract

Ground water validation is one of the most challenging issues facing modelers and hydrogeologists. Increased complexity in ground water models has created a gap between model predictions and the ability to validate or build confidence in predictions. Specific procedures and tests that can be easily adapted and applied to determine the validity of site-specific ground water models do not exist. This is true for both deterministic and stochastic models, with stochastic models posing the more difficult validation problem. The objective of this paper is to propose a general validation approach that addresses important issues recognized in previous validation studies, conferences, and symposia. The proposed method links the processes for building, calibrating, evaluating, and validating models in an iterative loop.

The approach focuses on using collected validation data to reduce uncertainty in the model and narrow the range of possible outcomes. This method is designed for stochastic numerical models utilizing Monte Carlo simulation approaches, but it can be easily adapted for deterministic models. The proposed methodology relies on the premise that absolute validity is not theoretically possible, nor is it a regulatory requirement. Rather, the proposed methodology highlights the importance of testing various aspects of the model and using diverse statistical tools for rigorous checking and confidence building in the model and its predictions. It is this confidence that will encourage regulators and the public to accept decisions based on the model predictions. This validation approach will be applied to a model, described in this paper, dealing with an underground nuclear test site in rural Nevada.

Introduction

During the past two decades, stochastic studies have shown that inadequate and insufficient data limit the ability of ground water models to predict system behavior without substantial uncertainty (Pohll et al. 1999; Pohlmann et al. 2000; Hassan et al. 2001). Uncertainty is always inherent in the model prediction and is the result of the inability to characterize fully the subsurface environment and the processes controlling the system behavior. Full characterization is limited by access to the subsurface, which requires extensive borehole drilling that can adversely affect the geologic integrity of the site or be prohibitively expensive.

Regulators and the public must accept modeling results in order to close subsurface-contaminated sites. Acceptance is difficult to secure, given the wide range of uncertainty associated with the predictions of stochastic models. A model validation process is probably the best way to address the acceptance issue as it can achieve buy-in for a closure process involving numerical ground water modeling. Validation, however, is not understood equally by all entities; there is an urgent need to unify the concepts of validation and develop a systematic way for testing and evaluating model predictions. A unified concept may facilitate acceptance of model-based decisions by regulators and the public, especially since many U.S. Department of Energy (DOE) and U.S. Department of Defense sites now undergoing closure processes require validation. Developing and using rigorous science to define a validation process that site sponsors, regulators, and the public can accept will be mutually beneficial.

The Central Nevada Test Area (CNTA), location of the Faultless underground nuclear test, is currently undergoing environmental restoration and facing the issue of

[1] Division of Hydrologic Sciences, Desert Research Institute, 755 E. Flamingo Road, Las Vegas, NV 89119; 702-862-5465; fax 702-862-5427; hassan@dri.edu (also at Irrigation and Hydraulics Department, Faculty of Engineering, Cairo University, Giza, Egypt)

Received August 2002, accepted April 2003.

model validation. Underground nuclear tests leave a significant radionuclide source in contact with ground water, without a technically feasible remediation technology. For these sites, regulatory closure and stewardship restrictions will depend on a model-generated contaminant boundary (perimeter of an area containing contamination exceeding a certain threshold). Confidence in the modeling results is critical in achieving closure. A complex, three-dimensional stochastic flow and transport model was developed for the CNTA site (Pohlmann et al. 2000). After reviewing the model, the state of Nevada determined that the model was acceptable for predicting contaminant boundaries for the CNTA, allowing a major step forward in the closure process (Chapman et al. 2002). Acceptance, however, was tied to a state requirement to validate the model. Thus, the CNTA model requires a validation strategy that can withstand the rigors of a scientific peer review, regulatory oversight, and public scrutiny.

Other sites urgently requiring effective validation strategies include Shoal underground test area in Nevada, DOE Hanford Site in Washington, Maxey Flats Nuclear Disposal Site in Kentucky, Fernald Environmental Project in Ohio, Oak Ridge National Laboratory in Tennessee, and Nevada Test Site in Nevada. Validation is of utmost concern to modelers, scientists, and regulatory agencies. A sound validation process requires procedures and tests, which do not currently exist, that can be easily adapted and applied to evaluate even the simplest deterministic model. Validation is even more difficult for predictive stochastic models that incorporate effects of parametric uncertainty and spatial variability.

A general methodology for validating numerical ground water models that addresses important issues acknowledged in previous validation studies, conferences, and symposia is presented in this paper. This method integrates various tools and strategies for evaluating predictive models, refining the predictions, reducing associated uncertainty, and building the confidence necessary to close sites having significant ground water contamination. The validation methodology focuses on steady-state models where the history matching concept advocated by Bredehoeft and Konikow (1993) is difficult to attain. Even if the system is under transient conditions, changes occur slowly, and it may not be feasible to establish a performance history to use in testing the model. Additionally, the proposed approach does not claim the model will be declared valid at some point in the process. That is, the approach deals with validation as a process, not an end result. The basic thrust is aimed at building confidence in the model to cover the crucial elements affecting the model predictions.

The remainder of the paper is organized as follows. Following this introduction, background information about the CNTA and the need for developing the validation methodology are discussed. The various definitions of validation and a review of previous validation studies are presented in the third section. Critical issues addressed in the validation process are presented in the fourth section, which builds upon previous ground water model validation studies and highlights important validation issues in these studies. The proposed validation strategy, with the necessary background information, is discussed in the fifth section. Conclusions are then presented. The statistical tools needed for the validation process are detailed in the appendices.

Description of the Central Nevada Test Area

CNTA is located approximately midway between Tonopah and Ely, Nevada (Figure 1). The study area is located within Hot Creek Valley, which extends ~110 km between north to south–oriented mountain ranges of the basin and range physiographic province. The valley is a long graben containing a thick sequence of Quaternary- and Tertiary-age fill (up to 1200 m) underlain and bounded on either side by Tertiary-age volcanic rocks (principally tuffs and rhyolite lavas). Annual precipitation averages 19.4 cm/year.

The water table at the Faultless site occurs almost 200 m below land surface (bls). Faultless was the only test performed at CNTA, and the working point is in the saturated zone at a depth of 975 m. Consequently, the major concern at the site is the transport of radioactive contaminants through the ground water system. This ground water system has two components—a shallow section (defined using data from less than 300 m bls) where flow is directed southward, and a deeper section (defined using data from 1500 to 2100 m bls) of regional flow directed northeastward. In the northern part of the valley, hydraulic head

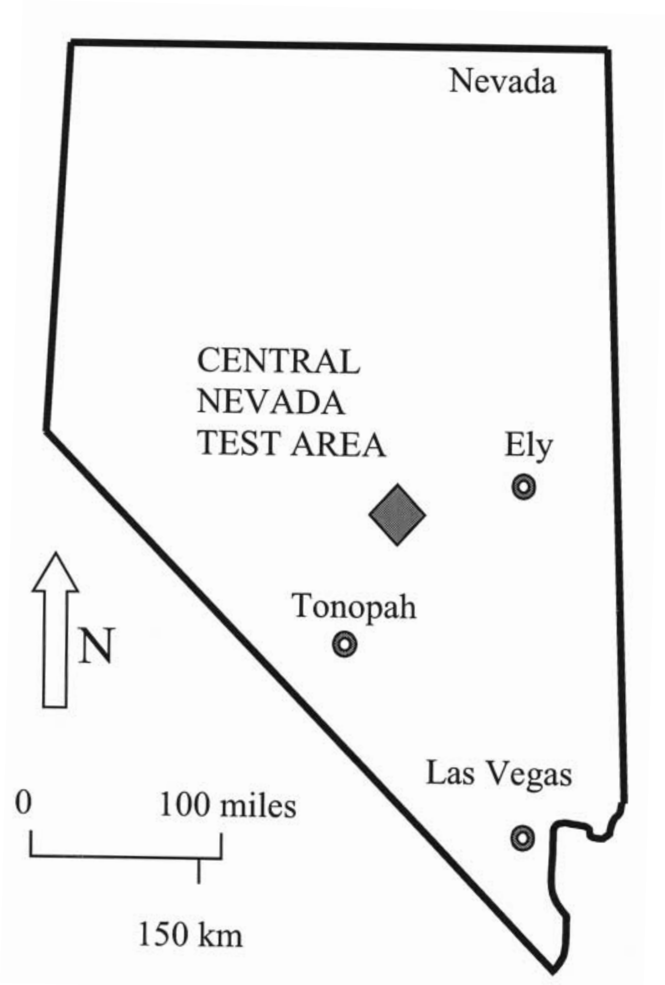

Figure 1. Map of Nevada showing the location of the Central Nevada Test Area.

decreases with increasing depth, indicating a recharging environment. In the southern part of the valley, head increases with increasing depth and artesian conditions are encountered, characteristic of a discharge area.

The conceptual flow model uses three principal hydrogeologic units including alluvium; tuffaceous sediments, bedded tuffs, and partially welded tuffs; and rhyolites and densely welded tuffs. The rhyolites and densely welded tuffs are assumed to be highly fractured and faulted and, where present, are considered to be the primary pathways for ground water flow and transport. Porous medium flow is assumed in the alluvium and tuffaceous sediments. The precise locations of the various units and respective values of hydraulic conductivity (*K*) are only known at the few borehole locations. The natural hydrogeologic heterogeneity is described in two aspects. The occurrence of the hydrogeologic units throughout the bulk of the model is allowed to be uncertain, and the assignment of heterogeneous *K* values to a unit is based on the variogram fitted to the available data for each of the three units.

Using the *K* maps as the foundation for ground water flow calculations, hundreds of equiprobable flow fields are created for the site. These in turn are the basis for transport calculations, performed using a random-walk, particle-tracking method. Complete details can be found in Pohlmann et al. (2000) and Chapman et al. (2002).

Though several aspects of uncertainty were included in the model, concerns remained regarding uncertainty in values of individual parameters. A data decision analysis (DDA) was performed (Pohll and Mihevc 2000) to quantify uncertainty in the existing model and determine the most cost-beneficial activities for reducing uncertainty, if necessary. The DDA indicated the overall uncertainty in the calculated contaminant boundary during the 1000-year regulatory time frame was relatively small, and only limited uncertainty reduction could be expected from expensive characterization activities. With these results, the model sponsor and the regulator determined the site model was suitable and the corrective action process could move forward. Key to this acceptance was the acknowledgment that the model requires independent validation data and requires long-term monitoring.

The proposed validation methodology is centered around three main themes. First is testing predictions of numerical ground water flow, transport models, and underlying conceptual models to determine if the assumptions are robust and consistent with regulatory purposes. Second is reevaluating and refining model predictions, and reducing the uncertainty based on data collected in the proposed field activities. Third is linking validation efforts to long-term monitoring efforts that benefit from, and build on, the validation-phase field activities. Though unique because of the nature of contamination, underground nuclear tests share much in common with other sources of ground water contamination; in particular, the problem of uncertainty. The proposed validation approach can therefore be adapted and applied to other sites of ground water contamination where predictive, site-specific models are used for decision-making purposes. The rigor of the proposed approach is based on its simplicity, comprehensiveness, and coverage of many aspects of the model, rather than its mathematical complexity.

Definitions of Model Validation and Review of Previous Studies

Definitions

Most controversies over the term validation arise from interpretations and perceived meanings. These range from an unachievable proof-of-truth view to more pragmatic approaches that emphasize subjective assessment in order to determine if models are good enough for a particular application (Zuidema 1994). Definitions can be generally grouped into four categories briefly summarized in the following paragraphs.

The first category relies on a scientific definition of validation, which defines validation as the demonstration that models are true representations of reality. The Nuclear Regulatory Commission defines validation as the process for ensuring that a model, as embodied in a computer code, is a correct representation of the intended process or system (NRC 1984). Anderson and Woessner (1992), Jackson et al. (1992), Oreskes et al. (1994), and others discussed similar views in more or less restrictive manners.

The second category is philosophical in nature and relies on the premise that a theory or hypothesis can never be validated, but only invalidated (Konikow and Bredehoeft 1992; Oreskes et al. 1994).

Various operational definitions constitute the third category views on model validation. Tsang (1987, 1991) describes the validation of a model with respect to a process or a site-specific system. In assessing the performance of nuclear repositories, McCombie et al. (1990) and Zuidema (1994) argue that a model is considered robust when there is confidence that errors will either have minimal effect on performance or yield conservative results.

The confidence-building views of model validation comprise the fourth category. For example, Neuman (1992) defines the validation of safety assessment models as the process of building scientific confidence in the methods used to perform these assessments. Additionally, Eisenberg et al. (1994) support the concept of confidence building. They indicate that this term acknowledges that full scientific validation of performance assessment models may be impossible, but models should be accepted based on appropriate testing to show that results are reasonable.

Previous Studies

International cooperative projects including INTRACOIN (1984), HYDROCOIN (Grundfelt et al. 1990), INTRAVAL (Nicholson 1990), and STRIPA (Herbert et al. 1990) focused on validating models. The subject was also extensively discussed in symposia including GEOVAL87 (1987), GEOVAL90 (1990), and GEOVAL94 (1994). The journal *Advances in Water Resources* dedicated two special issues to the topic of model validation (AWR 1992a, 1992b).

Most of these studies, international projects, and symposia focused on qualitative aspects of model validation.

Few touched on quantitative issues. In addition, some of the studies focused on validating a single aspect or observed phenomenon, e.g., matrix diffusion, and none addressed how to carry quantitatively a numerical, stochastic model through a validation process.

Critical Issues in the Model Validation Process

Reducing Prediction Uncertainty

Validating predictive models should provide confidence in the uncertainty bounds of the results where the real outcome will fall (Zuidema 1994). Although uncertainty cannot be eliminated, ways of making ground water models and subsequent decisions more reliable and effective are needed. The proposed plan focuses on using validation data to reduce uncertainty in the model and to narrow the range of possible outcomes of stochastic numerical models. The process will require iterative phases for collecting data, evaluating and refining the model, and reducing uncertainty in the model. This approach is particularly important in radionuclide transport models, as few aspects of the transport model results can actually be tested. In this case, the proposed validation approach would focus on the other elements of the model, e.g., geology, structure, flow, etc., and use the validation data to refine transport predictions and reduce uncertainty. These concepts should be clearly presented to model sponsors and regulators for their understanding and approval.

Diverse Data and Evaluation Tests

As discussed by Ababou et al. (1992), the degree to which a single experiment or a single set of field data can validate a model depends on subjective weights, or probability, assigned to the particular experiment. More validation weight can be assigned if the range of aspects covered by the experimental data set is sufficiently broad so the overall character of the model is efficiently tested. The field data should, therefore, be diverse and cover different aspects of the model. For example, the data should allow testing geologic aspects such as the existence and location of contact between different geologic units, flow model aspects such as head and gradient measurements, and transport or contaminant release aspects such as concentration measurements. Since a purpose of the validation task is to determine if multiple failures and far-field transport of contaminants can occur, transport aspects related to failure scenarios should be tested.

Oreskes et al. (1994) postulate that by using numerous and diverse confirming observations, it is reasonable to

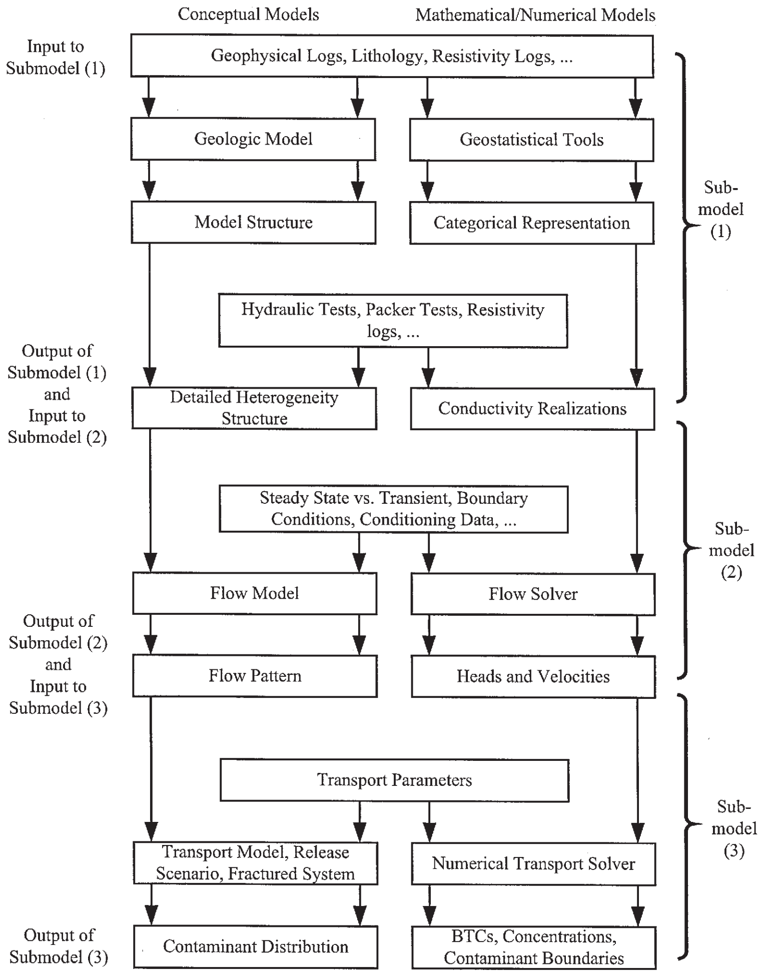
Conceptual Models
Mathematical/Numerical Models
Input to Submodel (1)
Geophysical Logs, Lithology, Resistivity Logs, ...
Geologic Model
Geostatistical Tools
Model Structure
Categorical Representation
Sub-model (1)
Hydraulic Tests, Packer Tests, Resistivity logs, ...
Output of Submodel (1) and Input to Submodel (2)
Detailed Heterogeneity Structure
Conductivity Realizations
Steady State vs. Transient, Boundary Conditions, Conditioning Data, ...
Sub-model (2)
Flow Model
Flow Solver
Output of Submodel (2) and Input to Submodel (3)
Flow Pattern
Heads and Velocities
Transport Parameters
Sub-model (3)
Transport Model, Release Scenario, Fractured System
Numerical Transport Solver
Output of Submodel (3)
Contaminant Distribution
BTCs, Concentrations, Contaminant Boundaries

Figure 2. Schematic of a general site-specific ground water model showing conceptual and numerical components, and three main submodels linked together.

conclude that the conceptualization embodied in the model is not flawed. A diversified set of statistical tests and evaluations will therefore provide a structured approach for evaluating the model predictions and building confidence in the decisions based on the predictions. The proposed systematic validation approach relies on tests and evaluation techniques to guide the decision regarding the model predictions, and support informed and grounded discussions among the modelers, sponsors, and regulators.

Submodels

In general, if a single model is composed of two or more submodels, the degree of confidence achieved by evaluating the submodels individually will not be as great as the degree of confidence achieved by evaluating the submodels linked together (Eisenberg et al. 1994). It is therefore important to conduct additional tests to validate the combined submodels. Site-specific ground water flow and transport models generally can be divided into three submodels that can be tested individually first and then in combination. Figure 2 is an example of the different submodels in a site-specific model and shows how they are linked to each other. The figure shows both conceptual and numerical submodels.

The first submodel is a conceptualized geologic model that identifies the different units and shows how they are structured together within the study domain. The input to the first submodel constitutes data types that help identify the geologic units and their locations, e.g., lithologic data, geophysical logs, resistivity logs, etc. Using geostatistical tools and conditional simulation with categorical or qualitative data, a discretized numerical submodel of the different categories or units can be obtained. Subsequently, the quantitative data, e.g., hydraulic testing results, packer tests, resistivity logs, etc., can be used to obtain the detailed heterogeneous structure of individual units in a quantitative manner. That is, the spatially varying hydraulic properties, namely hydraulic conductivity, can be obtained as an output of this first submodel.

For a general site-specific model and for the special case of the CNTA model, the first submodel can be tested in terms of the existence and location of the different units identified in the conceptual geologic model. Contact between the different units is also an important aspect that can be tested with validation data. For the CNTA model, Pohlmann et al. (2000) identify three geologic units with significant uncertainty associated with the contact between them. Conductivity values assigned to different layers should also be evaluated. This evaluation will focus on reducing uncertainty in the assigned conductivity values by utilizing head measurements and a conditional simulation (or inverse) approach. For example, the sequential self-calibration (SSC) approach (Gómez-Hernández et al. 1997) can be used for this purpose.

The second major submodel for a general site-specific ground water model is the flow submodel, where the output of submodel 1 is used as input. A conceptual flow model is then formulated and used in conjunction with this input, boundary conditions, and assumptions to derive the numerical flow model and solve the flow equations. This results in identifying the flow pattern in the simulation domain, which is represented by discretized head values and velocity components. This velocity distribution is the output of submodel 2 and is used as input to submodel 3.

The flow pattern at CNTA (and at many other field sites) is complicated (Pohlmann et al. 2000), and it is crucial to verify the directions of the vertical and lateral head gradients, especially in the vicinity of the contaminant source. Multiple head measurements at different levels can be obtained from a single borehole. These data will be crucial to testing the flow model and its underlying input data, as well as boundary conditions. In addition to testing the predicted heads, the head data will be used to reduce the heterogeneity uncertainty by using an inverse method such as the SSC approach.

In general, the last submodel in a site-specific study is the transport model. The conceptual transport model is identified by determining the source size and location, the release scenarios, and the transport processes encountered during the migration of contaminants. Added to the velocity pattern and boundary conditions, this conceptual model gives rise to the numerical transport model where the transport equations are formulated and solved for the output of concern. This solution yields temporal mass flux breakthrough curves at certain boundaries, spatial-temporal distribution of contaminant concentrations, or contaminant boundaries. Usually, these latter outputs are the target of the entire modeling process when ground water contamination is the major regulatory concern.

For the CNTA transport model, the release of radionuclides from the test cavity and the movement away from it are just beginning (based on a cavity infill time of 30 years for a test conducted in 1968). An important focus of validation of the transport aspects should be verifying the presence of fast migration channels, or failure scenarios that may have been overlooked and would thus lead to migration distances greater than the model predictions. Measurement of tritium concentrations in wells located sufficiently far from the cavity, i.e., beyond the fracturing radius to separate the possibility of fast migration pathways from prompt injection issues, will be important to test the adequacy of the transport model and whether the model (within its uncertainty bounds) has covered the critical transport issues.

After considering the different components and tests previously described, and linking the calibration analysis to the validation analysis, the linked submodels are evaluated. Flow of information between the three submodels provides a natural linkage that will enable collective evaluation of the entire model conducted in parallel with evaluations of individual submodels.

Subjective vs. Objective Judgment

Calculated and observed data for both calibration and validation processes are most often presented graphically with subjective interpretation of quality in the match (Flavelle 1992). It is generally preferred, however, to use a form of objective analysis in model calibration and validation. Objective quality is usually described by a goodness-of-fit parameter that reflects how well the model results match the observed calibration data. The goodness-of-fit parameter is usually used to optimize the calibration of the adjustable parameters in the model, and to serve as a measure for comparing alternate models. This is an inverse problem, where the main challenge is the nonuniqueness of the solution that yields different parametric values that provide solutions having similar accuracies (Poeter and Hill 1997; Hill et al. 1998; D'Agnese et al. 1999). The most common goodness-of-fit parameter appears to be a form of weighted root-mean-square error, with the error describing the difference between calculated and measured values. Unfortunately, the complexity of some evaluations makes them unattractive for general use by regulators and decision-makers (Flavelle 1992). Alternatively, simple goodness-of-fit tests can be used to describe the calibration and validation processes in an objective manner. The proposed validation approach relies heavily on objective evaluations, as well as a number of statistical measures and tests for evaluating different aspects of the model.

A common form of objective analysis for calibrating and validating simulation models is statistical hypothesis testing (Balci and Sargent 1981). This form of objective analysis can be used in addition to goodness-of-fit tests to evaluate the quality of the comparison between model predictions and measurements for both calibration and validation. Additional background information is presented in Appendix A (Goodness-of-Fit Measures) and Appendix B (Hypothesis Testing).

McCombie and McKinley (1993) argue the amount of effort dedicated to the validation process before the model is considered acceptable is necessarily subjective and depends on the complexity of the system and the initial objective for using the model. This argument highlights the fact that neither purely objective nor subjective judgment should be used exclusively in the validation process. Objective and subjective judgment are necessary components in the model validation process, and they complement each other. Model builders, model users, and regulators should agree that objective judgment would be complemented with subjective judgment and hydrogeologic expertise.

Validation Cost and Confidence in the Model

The cost of collecting data and conducting analyses should be considered in designing validation plans. As shown in Figure 3, adapted from Sargent (1990), there is a limit where increased investment does not significantly increase confidence in the model. The concerned parties must therefore agree on the level of confidence required for decision-making purposes, while considering the cost to achieve this level of confidence. The proposed validation approach, discussed in the next section, includes decision points regarding the cost of collecting the data and analyzing the model predictions.

Proposed Validation Approach

General

Mroczkowski et al. (1997) argue that validation using multi-response data is more powerful than traditional split-sample testing (a record of historical data is split into calibration and validation samples). Their argument, however, is based on validating conceptual catchment models where long historical records exist for the studied parameters. Using multi-response data could also be expected to be more powerful than single-response data in validating a subsurface flow and transport model. The proposed validation approach relies on both multi-response data and diverse statistical tests and analyses to evaluate model performance. By doing this, confidence can be built into model predictions, and field activities where data are collected for long-term monitoring can be guided.

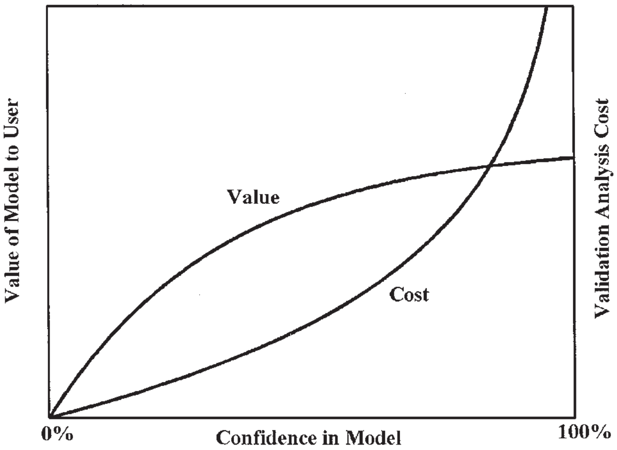

Figure 3. Value and validation cost of the model as functions of the desired level of confidence (adapted from Sargent [1990]).

To determine accuracy and adequacy of the model, the types and numbers of validation tests, degree of agreement between model and validation tests, and conformity between model descriptions and site-specific information should be considered (Davis et al. 1991, 1992). These authors emphasize rigorous development of the validation process and the importance of providing regulators with comprehensive information that follows a logical systematic approach. The proposed validation approach relies on numerous tests and evaluations, and follows a systematic approach. This approach, discussed in the next section, is particularly crucial in validating stochastic numerical models that rely on Monte Carlo simulation techniques where multiple realizations within this stochastic framework must be systematically analyzed and evaluated.

The CNTA validation plan is unique in that it is the first attempt to validate a stochastic model that explicitly accounts for spatial variability in conductivity and parametric uncertainty. The proposed validation process accounts for the stochastic nature of the model and attempts to reduce the realm of possibilities given by the large number of realizations considered in the Monte Carlo analysis.

Many of the tests proposed in the validation approach and their underlying principles are familiar. The power of these tests and the integrated validation approach stems from rigor and completeness, and not from innovation. New theories or statistical analyses are not being developed here, but rather available tools are being assembled to evaluate ground water models. These tests and the proposed validation method provide a structured approach for analyzing site-specific ground water models to build confidence in decisions based on the model predictions. Individual decisions throughout the validation stage will still be difficult, requiring subjective judgment and trade-offs, but using the proposed validation method will guide the decision and foster rational debate.

The philosophy underlying the development of the proposed validation approach relies on a forward-looking perspective. That is, by carrying the ground water modeling process beyond the iterative loop of characterization, calibration, modeling, prediction, and recharacterization to reduce uncertainty, much can be gained about the site and the model together. Unfortunately, regardless of the times the iterative process is repeated, uncertainty about the results of these studies will remain. Without a way to exit the iterative loop, resources may be wasted and the problem will remain unsolved. The flowchart shown in Figure 4 schematically represents this loop (steps 1–6) and proposes a logical way to exit the loop.

This exit occurs through the ground water flow and contaminant transport validation processes (the outer loop in Figure 4 comprising steps 7–9), which present a systematic method of determining when adequate confidence in the ground water model has been achieved and long-term monitoring should begin. It is possible, of course, that model deficiencies can drive the process back to the inner loop of characterization, but this would occur only after the validation and monitoring results are analyzed over time.

It is important to note that previous studies dealing with ground water model validation (Tsang 1991) focused only on the iterative loop shown in Figure 4. For example, Tsang (1991) asks if the evaluation of the results indicates the uncertainty is too large or if results with estimated uncertainty are good enough. This question correlates to step 5 in Figure 4. Additionally, previous studies have not explicitly considered the stochastic nature in a Monte Carlo fashion as considered in the proposed validation method. Furthermore, quantitative aspects were absent in previous studies, whereas the proposed method includes quantitative tools such as goodness-of-fit measures, hypothesis testing, and regression analysis.

Proceeding with validation analysis will produce additional data that may indicate the model is adequate, while

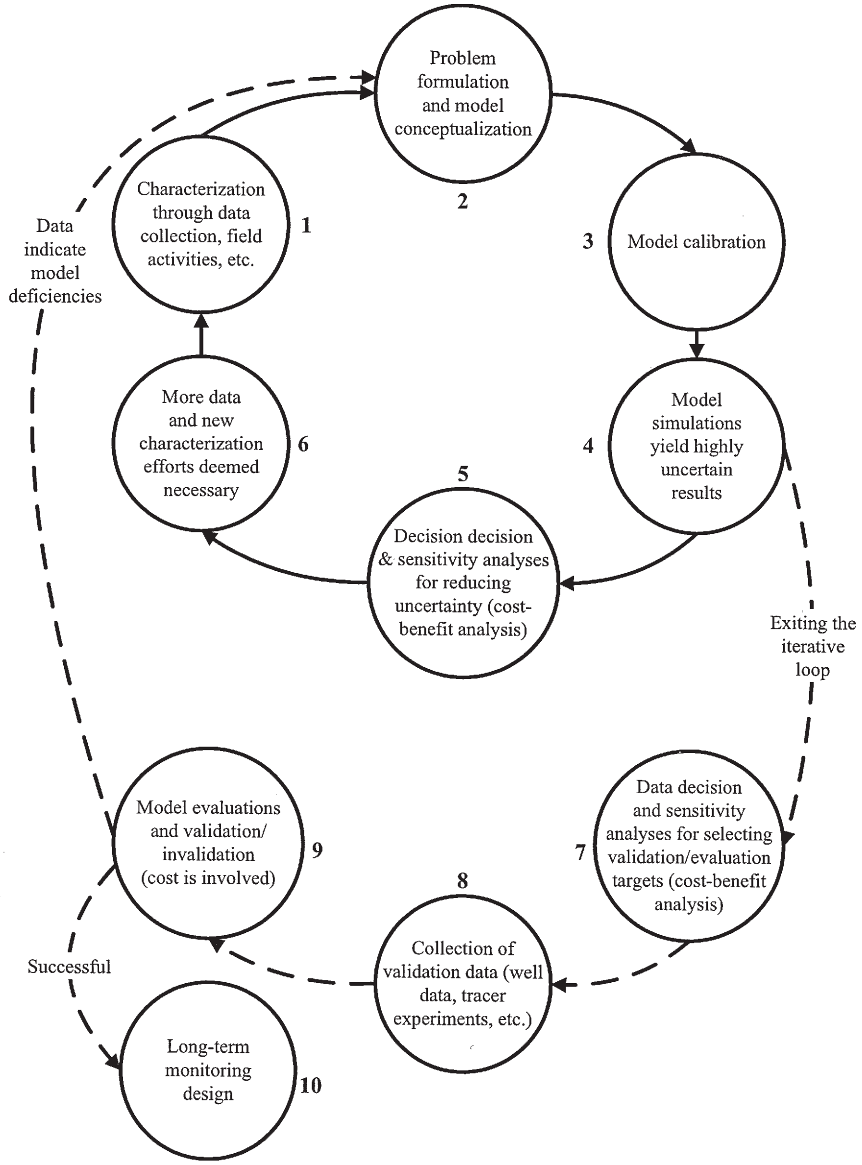
Problem formulation and model conceptualization
2
Characterization through data collection, field activities, etc.
1
Data indicate model deficiencies
3
Model calibration
4
Model simulations yield highly uncertain results
More data and new characterization efforts deemed necessary
6
5
Decision decision & sensitivity analyses for reducing uncertainty (cost-benefit analysis)
Exiting the iterative loop
Model evaluations and validation/ invalidation (cost is involved)
9
7
Data decision and sensitivity analyses for selecting validation/evaluation targets (cost-benefit analysis)
8
Collection of validation data (well data, tracer experiments, etc.)
Successful
Long-term monitoring design
10

Figure 4. Schematic of characterization, calibration, modeling, prediction, recharacterization iterative loop (steps 1–6) and decision to exit the loop, begin model validation process, and design long-term monitoring plan.

staying within the small iterative loop will not produce data to base a judgment regarding adequacy of the model. There will never be sufficient facts or data to eliminate all uncertainty or to produce a decision based solely on those facts. It is therefore better to move forward with uncertainty and evaluate how the model conforms to regulatory requirements, reevaluating decisions periodically.

Procedure for Proposed Model Validation Process

The eight procedural steps of the proposed model validation process are discussed in the following paragraphs. Detailed theoretical background and descriptions of the steps are presented in the appendices. The proposed steps are shown in the flowchart in Figure 5, which summarizes the steps and the iterative process for building confidence in ground water predictive models and to move toward long-term monitoring and closure of contaminated sites.

Step 1. Identify the data needed for validation, number and location of wells, and type of laboratory or field experiment needed. The well locations can be determined based on the existing model and should favor locations likely to encounter fast migration pathways. There are additional factors guiding well location that are determined by the site conditions and the nature of contamination. For example, in the CNTA model, the first consideration is that wells should be located far enough outside the fractured radius of the zone impacted by the nuclear test to avoid confusing prompt injection of radionuclides from the blast with radionuclide migration. Second, the wells should be located around the cavity in an orientation that will produce the most benefit in validating and refining the model. The layout of the wells around the contamination source should be designed to verify the lateral and vertical head gradients and flow directions. Other factors, such as safety associated with radioactive contamination and the cost of drilling and collecting data, have to be considered. Sequencing data collection is also important. Though it may be more practical and cost-efficient to drill the wells simultaneously, drilling one well at a time, collecting all possible data, and testing the model to determine the next field activity may be a better approach. Again, these choices depend on the specific problem and require a consensus among model developers and model users.

Step 2. Install the wells and collect as much data as possible from the wells. The data should include geophysical logging; resistivity logs; head measurements; concentrations, e.g., checking for tritium; and other information, e.g., temperature logs, conductivity measurements, that could be used to test the model structure, input, or output. The major portion of cost for deep ground water contamination, e.g., nuclear testing sites, is associated with drilling the wells. It is a good investment, therefore, to collect as much data as possible from the wells, because the extra cost for collecting additional data in the short-run will be marginal in comparison to drilling costs.

Step 3. Evaluate calibration accuracy for each realization using different goodness-of-fit measures in addition to the generalized likelihood uncertainty estimator (GLUE) (Freer et al. 1996; Franks and Beven 1997). This evaluation assumes that initially the model was qualitatively calibrated to minimize the deviation between model prediction and observed calibration data based mainly on visual inspection. A detailed discussion of the GLUE analysis is presented in Appendix C (Generalized Likelihood Uncertainty Estimate). Other tools, such as linear regression analysis, goodness-of-fit tests, and hypothesis testing, can be used to provide additional objective means to evaluate the relative strength of each realization in terms of reproducing the field calibration data. That relative strength will be linked later to the ability of individual realizations to match the validation data.

Step 4. Conduct validation tests to evaluate the components of the model and submodels. A promising stochastic validation approach was proposed by Luis and McLaughlin (1992) and was applied to a two-dimensional, deterministic, unsaturated flow model for predicting moisture movement during a field experiment conducted near Las Cruces, New Mexico. A detailed description of this approach is summarized from Luis and McLaughlin (1992) and presented in Appendix D (Stochastic Validation Approach). This approach can be adapted and used to test the flow model output (heads) under saturated conditions. Other objective tests, e.g., goodness-of-fit tests, can be used for the heads to complement this stochastic approach, which is based on hypothesis testing. Similar tests will be performed to test model structure and/or input depending on the type of data obtained in the field. Some data will be used to check the occurrence or absence of failure scenarios, e.g., at CNTA, verify if tritium exists farther from the cavity than predicted by any realization of the stochastic model. The intent is to evaluate individual realizations with as many diverse tests (in terms of the statistical nature of the test and the tested aspect of the model) as possible and to quantitatively measure the adequacy of each realization in capturing the main features of the modeled system.

Step 5. Link the results of the calibration accuracy evaluations and the validation tests for all realizations. Realizations can then be sorted based on adequacy and closeness to the field data. A subjective element may be invoked in the sorting process based on expert judgments and hydrogeologic understanding. The objective is to filter out the realizations that show a major deviation or inadequacy in the tested aspects and focus on realizations that passed the majority of the tests and evaluations. Consequently, the range of output uncertainty is reduced, and the subsequent effort can be focused on the most representative realizations and scenarios. To continue reducing the level of uncertainty, the conductivity distribution can be refined by using the SSC method described in Appendix E (Sequential Self-Calibration Method). In the SSC method, head (and concentration) measurements can be used to condition the generation of the conductivity field, so that the uncertainty in the conductivity heterogeneity pattern around each measurement location is reduced. The conductivity distribution for each original conductivity realization retained in the analysis can be updated.

Step 6. The results of step 5 will determine how the validation process continues. The number of realizations that attained a satisfactorily high score as compared to the number of realizations that attained a low score must be sufficient for further analysis.

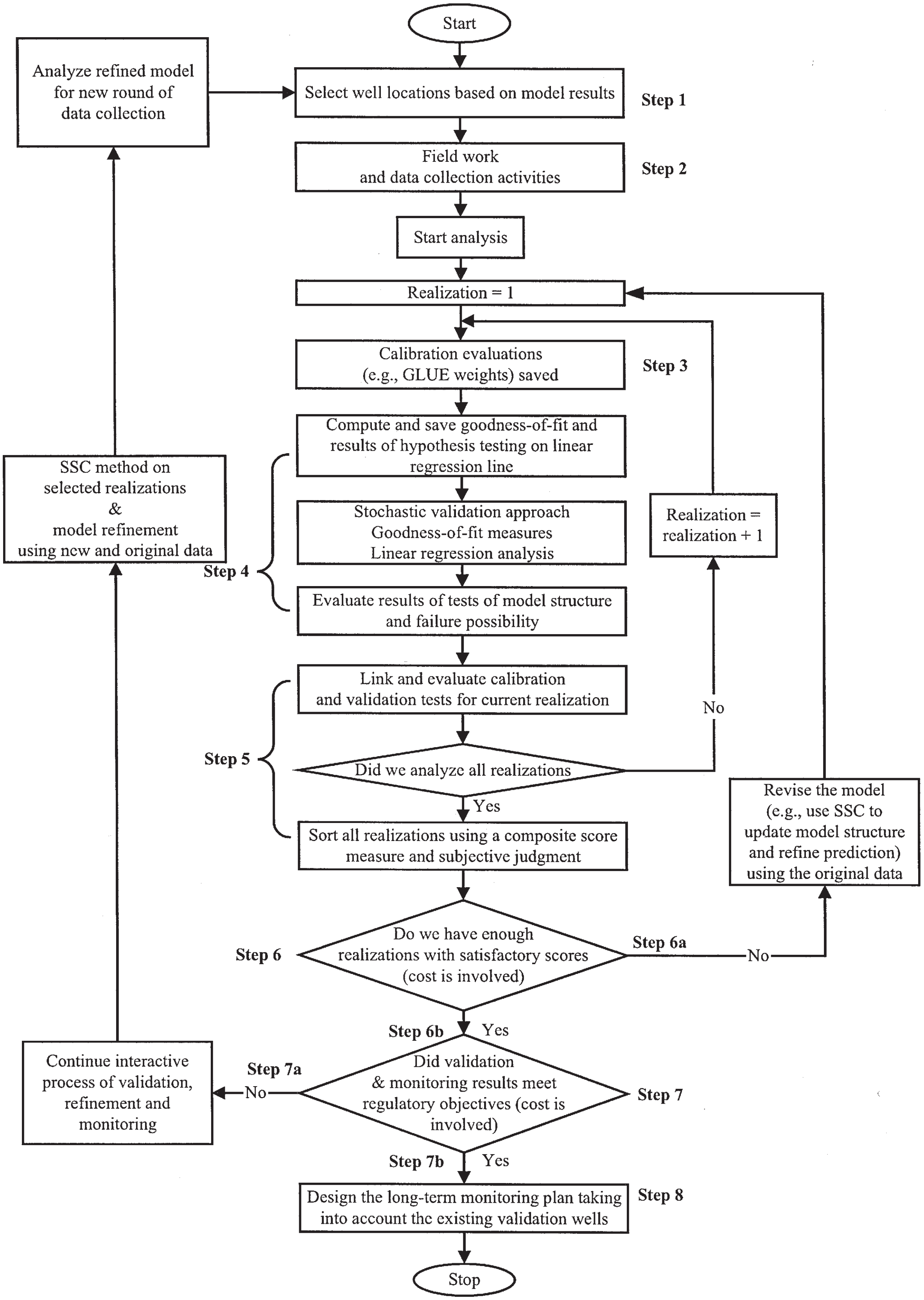
Start
Analyze refined model for new round of data collection
Select well locations based on model results
Step 1
Field work and data collection activities
Step 2
Start analysis
Realization = 1
Calibration evaluations (e.g., GLUE weights) saved
Step 3
Compute and save goodness-of-fit and results of hypothesis testing on linear regression line
SSC method on selected realizations & model refinement using new and original data
Stochastic validation approach Goodness-of-fit measures Linear regression analysis
Realization = realization + 1
Step 4
Evaluate results of tests of model structure and failure possibility
Link and evaluate calibration and validation tests for current realization
No
Step 5
Did we analyze all realizations
Yes
Revise the model (e.g., use SSC to update model structure and refine prediction) using the original data
Sort all realizations using a composite score measure and subjective judgment
Step 6
Do we have enough realizations with satisfactory scores (cost is involved)
Step 6a
No
Step 6b
Yes
Continue interactive process of validation, refinement and monitoring
Step 7a
No
Did validation & monitoring results meet regulatory objectives (cost is involved)
Step 7
Step 7b
Yes
Design the long-term monitoring plan taking into account the existing validation wells
Step 8
Stop

Figure 5. Flowchart of proposed validation approach and associated iterative refinement loops.

If the number of realizations having low scores is very large as compared to the total number of model realizations, the model probably has a major deficiency or conceptual problem or the input is incorrect. The conceptual model should be revised and the model structure updated based only on the original calibration data, if possible. The validation data should not be used. The validation data are set aside to avoid collecting new validation data when the previous analyses indicate the model is inadequate at this stage. If this is difficult, a compromise solution could be used where the validation data set is split and part of it is used in the model refinement process and the other part is saved for the next round of validation tests and analyses.

If the number of realizations having high scores is found sufficient, the model probably does not have major deficiencies or conceptual problems and the process can proceed forward.

Step 7. Once the rightmost loop in Figure 5 is completed successfully and a sufficient number of model realizations show acceptable performance (this decision will probably be based on the hydrologic expertise and judgment of the researchers), the model sponsors and regulators, in collaboration with the model developers, must determine whether the validation results meet the regulatory objectives. As suggested by Anderson and Woessner (1992), regulators should be content with some degree of partial validation and should further shift the focus from demands for validation to demands for a good modeling protocol that includes a complete description of the design, a thorough assessment of the calibration process, and an uncertainty analysis.

If more data are needed to build confidence in the model, then the decision loop on the left side in Figure 5 gives rise to a new iteration of model refinement, data collection, and re-evaluation. In this case, all available data become calibration data, and new validation data will probably need to be collected from new wells. Steps 1 through 6 will be repeated with the collection of new data based on the analysis of the refined model. The new wells for this round should be selected to serve two purposes—sources for new validation data and location targets for long-term monitoring.

If the answer to the question is yes, validation is deemed sufficient, the model is considered adequate or robust, and the process proceeds to Step 8.

Step 8. Design a long-term plan that includes setting and clarifying monitoring objectives; designing monitoring networks; and determining frequency, location, content, and schedule for sampling.

The preceding eight steps outline the proposed approach to use in validating stochastic numerical ground water models that rely on Monte Carlo simulations. The approach is general, and the application to the CNTA model may be the first attempt to validate a stochastic model for a nuclear testing site. The iterative nature of the proposed approach is one of its greatest strengths. Numerical ground water models and, in particular, stochastic models are very complex; modifying or changing any aspect may produce unanticipated consequences in another aspect of the model. To optimize the results of the validation process, one needs to separately consider the various details and take the broader view of the entire model while working systematically through the decisions and trade-offs.

Discussion

The process of validating a site-specific ground water model is difficult. It is not possible through the validation process to definitively confirm that the correct path is being followed. The author believes that confirmation is achieved from incremental information cumulatively collected through the various stages of the validation process. While it may not be possible to reach a conclusive outcome, the combined results and evaluations strongly improve the likelihood of an appropriate decision about model performance.

The proposed validation approach of site-specific ground water models is based on quantitative measures and statistical tests that are simple in design, yet diverse, so many aspects of the model can be tested. The approach must be implemented at an actual field site before it can be fully evaluated and all aspects analyzed. This validation approach is currently under discussion among the researchers who created the CNTA model (Pohlmann et al. 2000), the model sponsor (DOE), and the state regulator. The results will determine future direction in validating the model.

Results are essentially dictated by the conceptualization of the model. A portion of the proposed validation tests will be devoted to evaluating the conceptualization while other tests will be used to evaluate the results. For example, at CNTA, the conceptualization dictates that certain geologic layers exist in the modeled domain; the concept will be tested and verified through the validation data set. This conceptualization leads to the results, which show most radionuclide migration occurs in a particular geologic unit. If the validation data indicate the absence of the particular unit, then the conceptual model must be revised and results updated.

It is important to distinguish between model calibration and validation. Model calibration is a process whereby the model is tuned to identify the independent input parameters by fitting the model results to field or experimental data that usually represent the dependent system parameters. The calibration process can be quantitatively described by a goodness-of-fit measure. When the model is used to make long-term predictions, e.g., thousands of years at underground nuclear testing sites, the model is often calibrated using short-term data. Calibration cannot replace validation, and can only be considered as part of the site characterization and model formulation processes. In some situations, the validation task may become a calibration task whereby the experimental data collected for the validation purpose are used during the modeling effort. Although this type of calibration builds confidence in the model results, especially if the calibration fit is good, calibration by itself is not validation because the input parameters of the model are found based on the experimental results that can no longer be considered as validation data (Davis et al. 1991).

In the proposed validation approach, some of the validation data may eventually be used as calibration data, i.e., the rightmost and the leftmost loops in Figure 5. In each round of model validation or evaluation, however, a distinction is made between calibration data and associated calibration tests, i.e., the GLUE analysis, and validation data and associated validation tests, i.e., hypothesis testing and stochastic validation approach. These two independent sets of tests will be compared and linked together to derive the composite score that describes each realization of the model.

Conclusions

Common to most previous studies addressing ground water model validation is the absence of quantitative objective tools used in the proposed approaches, making it difficult to adapt the approach to different situations. Additionally, these studies present a consensus that absolute validity (accurate or exact representation of reality) is neither a theoretical possibility nor a regulatory requirement. Building confidence in the modeling process, and in the subsequent evaluation and validation process, is viewed as the best way to achieve model validation objectives and to instill acceptance by regulators and the public.

Building on previous studies, a ground water model validation strategy for stochastic numerical models is outlined. The proposed strategy accounts for important issues recognized in previous efforts. These issues include reducing prediction uncertainty, increasing diversity of data and evaluation tests, relying on objective measures whenever possible and capitalizing on subjective judgment and hydrogeologic insights, testing submodels separately and jointly, and recognizing that the cost element of the validation process will significantly impact decisions throughout the process. Consideration of these issues, and the fact that the confidence-building process is a long-term and iterative process, yield a systematic approach for the general case of a stochastic numerical model that relies on Monte Carlo simulations.

One of the main objectives of this study was to develop an integrated validation approach that relies on an iterative calibration-modeling-monitoring-evaluation-refinement cycle, thereby increasing confidence in the model predictions and reducing the level of uncertainty. This proposed validation approach will be implemented for a ground water flow and transport model of an underground nuclear testing site located at CNTA. The model has been accepted by the state of Nevada regulators with one condition—validating the model predictions. The validation methodology proposed in this paper will be fully developed, tested, and enhanced during the implementation and application to the CNTA underground nuclear testing site.

Acknowledgments

The author would like to thank Jenny Chapman, Greg Pohll, and Karl Pohlmann of the Desert Research Institute for stimulating discussion and comments. Thanks are also extended to Marjory Jones and Jennifer Lease for their editorial comments. In addition, the author is very thankful to John Bredehoeft, Carl Nuzman, Tiraz Birdie, and other anonymous reviewers for their constructive comments. Thanks are extended to Mary Anderson, editor in chief of *Ground Water*, for helpful comments and suggestions. This work was performed under the auspices of the U.S. Department of Energy, National Nuclear Security Administration Nevada Site Office by the Desert Research Institute under Contract Number DE-AC08–00NV13609.

Appendices: Background and Theoretical Concepts

This background information is compiled from the cited studies. The tools discussed within the appendices are just examples of many statistical techniques that may be used to achieve similar goals of the proposed validation methodology.

Appendix A: Goodness-of-Fit Measures

Legates and McCabe (1999) argue that correlation and correlation-based measures, e.g., coefficient of determination R^2, are oversensitive to extreme values or outliers, and insensitive to additive and proportional differences between model predictions and observations. They conclude that additional evaluations such as summary statistics and absolute error measures should supplement these goodness-of-fit measures to evaluate the model. They also present useful alternative goodness-of-fit and relative error measures, e.g., coefficient of efficiency, index of agreement, that overcome many of the limitations of correlation-based measures. The remainder of this appendix is a summary of the presentation of Legates and McCabe (1999) relevant to model evaluation tools.

Coefficient of Determination R^2

The coefficient of determination, R^2, describes the proportion of the total variance in the observed data that can be explained by the model and ranges from 0.0 to 1.0, with higher values indicating better agreement:

$$R^2 = \frac{\sum_{i=1}^{N} (O_i - \overline{O})(P_i - \overline{P})}{\left[\sum_{i=1}^{N} (O_i - \overline{O})^2\right]^{0.5} \left[\sum_{i=1}^{N} (P_i - \overline{P})^2\right]^{0.5}} \qquad \text{(A1)}$$

where the overbar denotes the mean, P denotes predicted variable, O indicates observed values, and N is the number of available pairs of predicted vs. measured values. If $P_i = (AO_i + B)$ for any nonzero value of A and any value of B, then $R^2 = 1.0$. Thus, R^2 is insensitive to additive and proportional differences between the model predictions and observations. It is also more sensitive to outliers than to observations near the mean.

Coefficient of Efficiency E

The coefficient of efficiency, which ranges from minus infinity to 1.0, is defined as

$$E = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} (O_i - \overline{O})^2} \qquad \text{(A2)}$$

The coefficient of efficiency represents an improvement over R^2 for model evaluation purposes in that it is sensitive to differences in the observed and model-simulated means and variances; that is, if $P_i = (AO_i + B)$, then E decreases as A and B vary from 1.0 and 0.0, respectively. Because of the squared differences, however, E is overly sensitive to extreme values as is R^2.

Index of Agreement d

The index of agreement, d, was developed to overcome the insensitivity of correlation-based measures to additive and proportional differences between observations and model simulations. It is expressed as

$$d = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} (|P_i - \overline{O}| + |O_i - \overline{O}|)^2} = 1 - N\,\frac{\text{MSE}}{\text{PE}} \tag{A3}$$

The index of agreement varies from 0.0 to 1.0 and represents an improvement over R^2, but it is also sensitive to extreme values owing to the squared differences.

The sensitivity of R^2, E, and d to extreme values led to the suggestion that a more generic index of agreement could be used in the form

$$d_j = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^j}{\sum_{i=1}^{N} (|P_i - \overline{O}| + |O_i - \overline{O}|)^j} \tag{A4}$$

where j represents an arbitrary power, i.e., a positive integer. The original index of agreement d given in Equation A3 becomes d_2 using this notation. For $j = 1$, the resulting index, d_1, has the advantage that errors and differences are given their appropriate weighting, not inflated by their squared values. Similarly, the coefficient of efficiency can be adjusted as

$$E_j = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^j}{\sum_{i=1}^{N} (O_i - \overline{O})^j} \tag{A5}$$

In addition to E and d measures, absolute error measures should be considered, which include the root mean square error ($\text{RMSE} = \sqrt{\text{MSE}}$) and the mean absolute error$\left(\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |O_i - P_i|\right)$. These additional measures describe the differences between the model simulations and observations in the units of the predicted variable.

The analyses of Legates and McCabe (1999) were based on analyses of time-series models where large data sets are available to test prediction models. In the subsurface, however, availability of such abundant data is not common. Some of the preceding measures, therefore, may not be usable for such limited data. It is thus important to use as many measures as possible to get a better evaluation of the model predictions.

Appendix B: Hypothesis Testing

Statistical hypothesis testing can be used as a quantitative tool for evaluating predictive models. The test usually postulates a null hypothesis (H_0) and a complementary hypothesis (H_1). The null hypothesis postulates the assumption or result that requires testing, e.g., the model is valid or the linear regression line has a slope of unity, while the complementary hypothesis postulates the opposite. Two types of errors can occur in hypothesis testing with certain probabilities—Type I errors and Type II errors. The probability of Type I error is called the model builder's risk (α), whereas the probability of Type II error is called the model user's risk (β). In model validation, the model user's risk is extremely important and must be kept small (Sargent 1990). These probabilities and those for making the right decisions are shown in Table 1, adapted from Balci and Sargent (1981). Both Type I and Type II errors must be considered in using hypothesis testing for model validation, and the risks resulting from these errors can be decreased at the expense of increasing the sample sizes of observations.

There is a direct relation among model builder's risk, model user's risk, acceptable validity range (amount of acceptable accuracy required for the model to be valid under a given set of experimental conditions), and sample size of observations (equivalent to a cost parameter). The model sponsor, model user, and model builder for the intended application of the model can make a trade-off among these parameters (Balci and Sargent 1981).

Appendix C: Generalized Likelihood Uncertainty Estimate

To honor site-specific data during calibration and subsequent modeling, the generalized likelihood uncertainty estimator (GLUE) algorithm can be used (Freer et al. 1996; Franks and Beven 1997). The GLUE procedure is an extension of Monte Carlo random sampling to incorporate the goodness-of-fit measure for each simulation. A likelihood measure is an evaluation of the quantitative goodness-of-fit. For example, the likelihood estimator for the solution of the flow equation can be defined as

$$L(Y|\Theta) = \left[\sum (\varepsilon)^2\right]^{-M} \tag{C1}$$

where

$$\varepsilon = h_j^* - \hat{h}_j \tag{C2}$$

Table 1
Outcomes of Hypothesis Testing

Result of Hypothesis Testing	Actual Status of Model	
	Model Is Valid Null Hypotheis, H_0 Is True	**Model Is Invalid Complementary Hypothesis, H_1 Is True**
Do not reject H_0	Correct decision	Model user's risk β
Reject H_0	Model builder's risk α	Correct decision

(Adapted from Balci and Sargent [1981])

and $L(Y|\Theta)$s the likelihood of the vector of outputs, Y, knowing Θ, the vector of random inputs; $\hat{h}_j$ the simulated head at the point j; h_j^* is the observed head at that point; and M is a likelihood shape factor. Although the choice of the M factor is subjective, its value defines its relative function. As M approaches zero, likelihood approaches unity, and each simulation has equal weight as is the case with traditional Monte Carlo analysis. As M approaches infinity, simulations with the lowest sum of squared errors (the simulations that best fit the field data) receive essentially all the weight, which is analogous to an inverse solution. The likelihood weights calculated for each realization based on Equation C1 can be used in subsequent modeling to give more weight to those realizations that best fit the field data during the calibration process. Additionally, these weights can be used later in the validation stage to compare the performance of individual realizations when acquiring new field data for validation analysis.

Appendix D: Stochastic Validation Approach (Luis and McLaughlin 1992)

Luis and McLaughlin (1992) proposed and applied a stochastic approach that relied on hypothesis testing to validate a two-dimensional, deterministic, unsaturated flow model for predicting moisture movement at a field site near Las Cruces, New Mexico. The approach began by identifying the factors that contributed to the differences between model predictions and observations (for simplicity, the predicted parameter is assumed to be the hydraulic head in a saturated system).

These differences were attributed to the following three sources of error. First are measurement errors, which represented the differences between the true values and the small-scale values of the hydraulic head. Second is the spatial heterogeneity, which represented the difference between the large-scale trend (or smoothed head) that the model was intended to predict and the true small-scale values of head. Third is model error, which represented the difference between the model prediction and the actual smoothed trend. Figure D1a shows schematic representations of these error sources where an actual, *hj*, fluctuating (due to heterogeneity) head distribution with a large-scale trend, $\bar{h}_j$, is shown in conjunction with a hypothesized stepwise distribution representing model prediction, $\hat{h}_j$. Measurement errors are only dependent on the measurement protocol and accuracy of the device that is used, which are not related to the model. The spatial heterogeneity effect is embedded in the difference between the small-scale measurements and the large-scale trend, and this difference is not really an error, but a reflection of the difference in scale between the measured and predicted quantities (Luis and McLaughlin 1992). Model error is a reflection of the ability of the model to predict the large-scale trend, which is the primary quantity of interest in this case and could be due to conceptual deficiencies or erroneous inputs.

The j^{th} measurement residual, ε_j, observed at location $\mathbf{x}_j$ (for $j = 1, \ldots N$) where N is the total number of head measurements used for validation can be written as

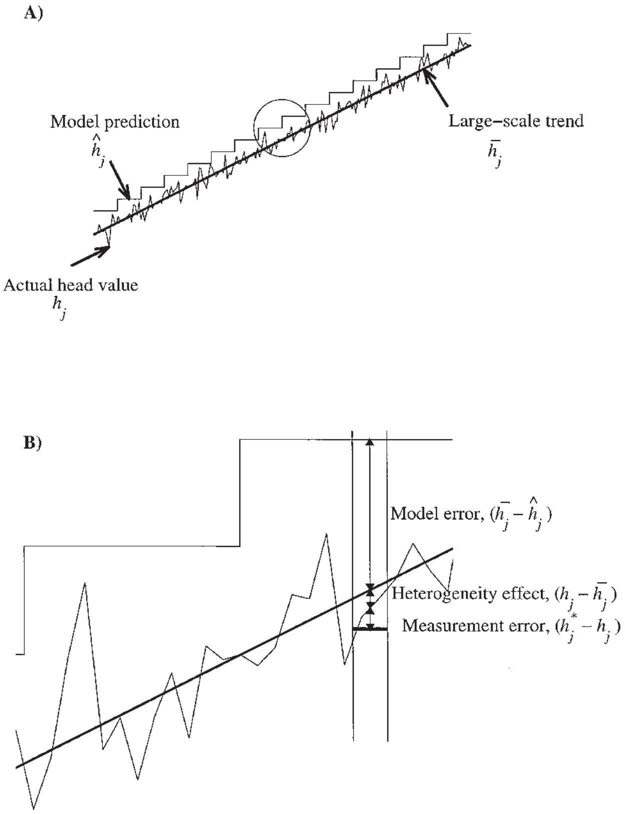
A)
Model prediction
$\hat{h}_j$
Large–scale trend
$\bar{h}_j$
Actual head value
h_j
B)
Model error, $(\bar{h}_j - \hat{h}_j)$
Heterogeneity effect, $(h_j - \bar{h}_j)$
Measurement error, $(h_j^* - h_j)$

Figure D1. (a) Schematic of actual head distribution, large-scale trend, and stepwise model prediction, and (b) decomposition of measurement residual into three error components.

$$\varepsilon(\mathbf{x}_j) = \varepsilon_j = h_j^* - h_j^{\wedge} - (\hat{\eta})$$

$$= [h_j^* - h_j] + [h_j - \bar{h}_j] + [\bar{h}_j - \hat{h}_j(\hat{\eta})] \quad \text{(D1)}$$

where $h_j^* = h^*(\mathbf{x}_j)$ is the head measurement at $\mathbf{x}_j$, and $\hat{h}_j = \hat{h}$ $(\mathbf{x}_j|\hat{\eta})$ is the model prediction at the same location obtained by using a set of estimated model parameters, $\hat{\eta}$, $h_j = h(\mathbf{x}_j)$ is the true head value at $\mathbf{x}_j$, and $\bar{h}_j = \bar{h}(\mathbf{x}_j)$ is the smoothed value of the large-scale trend or the expected value of h_j. The first term between the square brackets in Equation D1 represents measurement error, the second bracketed term represents the effect of geologic heterogeneity, and the last term represents the model error. These errors are schematically shown in Figure D1b. It is assumed here that the mathematical expectation of the head represents the large-scale, e.g., at the 50 m grid scale of the CNTA model, values of head that govern the flow pattern and the transport velocities.

If the model prediction is equal to the smoothed, large-scale values, the model error term in Equation D1 is zero. In statistical terms, the following null hypothesis is considered:

$$\text{H}_0\text{: Model error is negligible, } \hat{h}_j(\hat{\eta}) = \bar{h}_j$$

$$\text{H}_1\text{: Model error is significant, } \hat{h}_j(\hat{\eta}) \neq \bar{h}_j \quad \text{(D2)}$$

To apply this hypothesis-testing technique to the model validation problem, one must find test statistics that can be used to check the hypothesis defined in Equation D2. These statistics should depend on available head measurements and should be designed to minimize the risk associated with making erroneous decisions about hypothesis testing (see Appendix B on hypothesis testing and associated errors). If one designs a stringent test, the model user's risk, β, will be small, and the model builder's risk, α, will be large (tending incorrectly to reject acceptable models). If, on the other hand, the test is less stringent, it will have a large β and a small α (tending incorrectly to accept defective models).

When the hypothesis in Equation D2 is true, the variance of the measurement residual, $\sigma^2_{\varepsilon_j}$, can be written as (Luis and McLaughlin 1992)

$$\sigma^2_{\varepsilon_j} = \sigma^2_{h^*} + \sigma^2_{h_j} \tag{D3}$$

where $\sigma^2_{h^*}$ is the measurement error variance and $\sigma^2_{h_j}$is the head variance. This head variance, $\sigma^2_{h_j}$, plays a key role in this approach since it defines the amount of variability one should expect in the model's predictions when the model structure and measurements are both perfect. In other words, this variance establishes a type of lower bound on the ability of the model to predict point values of head (Luis and McLaughlin 1992). If the head variance can be derived directly from the numerical results of the flow model, e.g., using Monte Carlo simulations, Equation D3 can be used to evaluate the measurement residual variance to be expected when hypothesis H_0 in Equation D2 is true.

If the actual residual variance is much larger, it can be presumed that H_0 is not true, i.e., model errors are significant. Luis and McLaughlin (1992) suggest a number of tests that focus on testing whether the mean residual is zero, testing whether the mean squared residual is smaller than a certain tolerance, and analyzing the spatial structure of the residuals. These tests can be applied to all available measurements or to selected subsets.

In their application to the Las Cruces experiment, which has an unusually extensive set of soil data and validation measurements collected over horizontal and vertical distances of several meters and over time scales of a few years, Luis and McLaughlin (1992) could not reach a conclusion regarding the ability of the model to predict the observed moisture content at later times. In addition, Ababou et al. (1992) assert that this approach, although very valuable, is not complete since the hypothesis that the model is false remains untested, and the probability of accepting a false model cannot be evaluated by this technique. To evaluate this possibility, one would need to postulate another complementary model, or class of models, known to be always true if the model is false.

This critique of the Luis and McLaughlin (1992) approach and the incompleteness of hypothesis-testing techniques emphasize the need for conducting as many tests as possible to evaluate model performance. Since these statistical tests are not exact, it is beneficial to consider all tests together and link the calibration results to the results of the validation tests for each individual realization as was shown in Figure 5. Although the possibility exists theoretically, it is highly unlikely for an individual realization to pass the majority of tests and represent a false model. If this realization is accepted as valid based on the results of numerous tests, it is reasonable to assume that the model user's risk, β, is very small. On the other hand, if an individual realization fails to pass a large number of the tests, then rejecting this realization as invalid is not expected to represent a large Type I error.

Appendix E: Sequential Self-Calibration Method

To continue reducing the uncertainty level, a refinement in conductivity distribution can be made using the sequential self-calibration (SSC) method. In this method, new head measurements (and old ones) can be used to condition the generation of the conductivity field in a way that the uncertainty in the conductivity heterogeneity pattern around each measurement location is reduced.

Several new geostatistically based inverse approaches have been developed to generate the hydraulic conductivity fields by conditioning on both the hydraulic head and conductivity measurements (Zimmerman et al. 1998). Among these new approaches, the SSC method (Gómez-Hernández et al. 1997; Capilla et al. 1998; Wen et al. 1999) is an iterative, geostatistically based inverse technique that allows generation of multiple, equiprobable realizations of heterogeneous fields that match the dynamic data in addition to the typical geostatistical constraints. In the validation process, we can use this method in the refinement portion of the iterative loop of modeling, validation, and refinement. This, or similar methods, can be systematically used to refine the model predictions and reduce their uncertainty bounds.

The main steps in the SSC method are adapted and summarized here within the application to the validation approach. The process begins with the original hydraulic conductivity fields that were generated for the model to be evaluated and validated. Using the flow and transport solutions provided by the original model for individual realizations, the realizations are processed one at a time using the new (as well as old) data collected for validation purposes. An objective function that measures the mismatch between predicted and observed head and concentration data can then be written as (Wen et al. 1999)

$$O = \sum_{1}^{nwell} W_c(nw)[\hat{C}(nw) - C(nw)]^2 + \sum_{1}^{nwell} W_h(nw)[\hat{h}(nw) - h(nw)]^2 \tag{E1}$$

where $W_c(nw), W_h(nw)$ are the weights assigned to the concentration and head-sampling well, *nw*, according to sampling accuracy. Matching the head and concentration data is achieved by minimizing the objective function. A gradient-based method is used for optimization, which requires calculating sensitivity coefficients, the derivatives of concentration and head with respect to the hydraulic conductivity perturbation:

$$\frac{\partial \hat{C}(nw)}{\partial \Delta K_i}, \frac{\partial \hat{h}(nw)}{\partial \Delta K_i} \quad i = 1,\cdots,N \tag{E2}$$

where N is the number of blocks in the model. The optimal changes of conductivity are determined at selected master points (Gómez-Hernández et al. 1997) and then smoothly interpolated by kriging to all grid blocks. One would then go back and evaluate the objective function until it is sufficiently close to zero, or less than a predetermined tolerance value. Fewer than 20 iterations are normally required (Wen et al. 1999).

References

Ababou, R., B. Sagar, and G. Wittmeyer. 1992. Testing procedures for spatially distributed flow models. *Advances in Water Resources* 15, 181–198.

Anderson, M.P., and W.W. Woessner. 1992. The role of postaudit in model validation. *Advances in Water Resources* 15, 167–173.

AWR. 1992a. Special issue: Validation of geo-hydrological models, Part 1. *Advances in Water Resources* 15, no. 1.

AWR. 1992b. Special issue: Validation of geo-hydrological models, Part 2. *Advances in Water Resources* 15, no. 3.

Balci, O., and R.G. Sargent. 1981. A methodology for cost-risk analysis in the statistical validation of simulation models. *Communication of the ACM* 24, no. 4: 190–197.

Bredehoeft, J.D., and L.F. Konikow. 1993. Ground water models: Validate or invalidate. *Ground Water* 31, no. 2: 178–179.

Capilla J., J. Gómez-Hernández, and A. Sahuquillo. 1998. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, III: Application to the Culebra formation at the Waste Isolation Pilot Plant (WIPP), New Mexico, USA. *Journal of Hydrology* 207, no. 3–4: 254–269.

Chapman, J.B., K. Pohlmann, G. Pohll, A.E. Hassan, P. Sanders, M. Sanchez, and S. Jaunarajs. 2002. Remediation of the Faultless underground nuclear test: Moving forward in the face of model uncertainty. In *Proceedings of the Waste Management '02 Conference*, Tucson, Arizona. Tucson, Arizona: WM Symposia Inc.

D'Agnese, F.A., C.C. Faunt, M.C. Hill, and A.K. Turner. 1999. Death Valley regional ground water flow model calibration using optimal parameter estimation methods and geoscientific information systems. *Advances in Water Resources* 22, no. 8: 777–790.

Davis, P.A., N.E. Olague, and M.T. Goodrich. 1991. Approaches for the validation of models used for performance assessment of high-level radioactive waste repositories. (Sandia National Laboratories, SAND90–0575, Albuquerque, New Mexico). Washington, D.C.: Division of High Level Waste Management, Office of Nuclear Material Safety and Safeguards, U.S. Nuclear Regulatory Commission.

Davis, P.A., N.E. Olague, and M.T. Goodrich. 1992. Application of a validation strategy to Darcy's experiment. *Advances in Water Resources* 15, 175–180.

Eisenberg, N., M. Federline, B. Sagar, G. Wittmeyer, J. Andersson, and S. Wingefors. 1994. Model validation from a regulatory perspective: A summary. In *Proceedings GEOVAL '94: Validation Through Model Testing*, NEA/SKI Symposium, Paris, France, 421–434. Paris: Nuclear Energy Agency.

Flavelle, P. 1992. A quantitative measure of model validation and its potential use for regulatory purposes. *Advances in Water Resources* 15, 5–13.

Franks, S.W., and K.J. Beven. 1997. Bayesian estimation of uncertainty in land surface-atmosphere flux predictions. *Geophysical Research* 102, no. D20: 23991–23999.

Freer, J., K. Beven, and B. Ambroise. 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research* 32, no. 7: 2161–2173.

GEOVAL87. 1987. In *Proceedings GEOVAL '87: Verification and Validation of Geosphere Performance Assessment Models*, SKI Symposium, Stockholm, Sweden. Stockholm: Swedish Nuclear Power Inspectorate.

GEOVAL90. 1990. In *Proceedings GEOVAL '90: Validation of Geosphere Flow and Transport Models,* SKI Symposium, Stockholm, Sweden. Stockholm: Swedish Nuclear Power Inspectorate.

GEOVAL94. 1994. In *Proceedings GEOVAL '94: Validation Through Model Testing*, NEA/SKI Symposium, Paris, France. Paris: Nuclear Energy Agency.

Gómez-Hernández, J.J., A. Sahuquillo, and J.E. Capilla. 1997. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, I: The theory. *Journal of Hydrology* 203, no. 1–4: 162–174.

Grundfelt, B., B. Lindbom, A. Larsson, and K. Andersson. 1990. HYDROCOIN level 3: Testing methods for sensitivity/uncertainty analysis. In *Proceedings GEOVAL '90: Validation of Geosphere Flow and Transport Models,* SKI Symposium, Stockholm, Sweden. Stockholm: Swedish Nuclear Power Inspectorate .

Hassan, A.E., K.F. Pohlmann, and J.B. Champan. 2001. Uncertainty analysis of radionuclide transport in a fractured coastal aquifer with geothermal effects. *Transport in Porous Media* 43, 107–136.

Herbert, A., W. Dershowitz, J. Long, and D. Hodgkinson. 1990. Validation of fracture flow models in the Stripa project. In *Proceedings GEOVAL '90: Validation of Geosphere Flow and Transport Models,* SKI Symposium, Stockholm, Sweden. Stockholm: Swedish Nuclear Power Inspectorate.

Hill, M.C., R.L. Cooley, and D.W. Pollock. 1998. A controlled experiment in ground water flow model calibration. *Ground Water* 36, no. 3: 520–535.

INTRACOIN. 1984. Final report level 1: Code verification. Report SKI 84:3. Stockholm: Swedish Nuclear Power Inspectorate.

Jackson, C.P., D.A. Lever, and P.J. Summer. 1992. Validation of transport models for use in repository performance assessment: A view illustrated for INTRAVAL test case 1b. *Advances in Water Resources* 15, 33–45.

Konikow, L.F., and J.D. Bredehoeft. 1992. Ground water models cannot be validated. *Advances in Water Resources* 15, 75–83.

Legates, D.R., and G.J. McCabe Jr. 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35, no. 1: 233–241.

Luis, S.J., and D. McLaughlin. 1992. A stochastic approach to model validation. *Advances in Water Resources* 15, 15–32.

McCombie, C., I.G. McKinley, and P. Zuidema. 1990. Sufficient validation: The value of robustness in performance assessment and system design. In *Proceedings GEOVAL '90: Validation of Geosphere Flow and Transport Models,* SKI Symposium, Stockholm, Sweden, 598–610. Stockholm: Swedish Nuclear Power Inspectorate.

McCombie, C., and I. McKinley. 1993. Validation—Another perspective. *Ground Water* 31, no. 4: 530–531.

Mroczkowski, M., G.P. Raper, and G. Kuczera. 1997. The quest for more powerful validation of conceptual catchment models. *Water Resources Research* 33, no. 10: 2325–2335.

Neuman, S.P. 1992. Validation of safety assessment models as a process of scientific and public confidence building. In *Proceedings of High Level Waste Management Conference*, Vol. 2, Las Vegas. LaGrange Park, Illinois: American Nuclear Society.

Nicholson, T.J. 1990. Recent accomplishments in the INTRAVAL Project—A status report on validation efforts. In *Proceedings GEOVAL '90: Validation of Geosphere Flow and Transport Models,* SKI Symposium, Stockholm, Sweden. Stockholm: Swedish Nuclear Power Inspectorate.

NRC. 1984. A revised modeling strategy document for high-level waste performance assessment. Washington, D.C.: U.S. Nuclear Regulatory Commission.

Oreskes, N., K. Shrader-Frechette, and K. Belits. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 264, 641–646.

Poeter, E.P., and M.C. Hill. 1997. Inverse models: A necessary next step in ground water modeling. *Ground Water* 35, no. 2: 250–260.

Pohll, G., A.E. Hassan, J.B. Chapman, C. Papelis, and R. Andricevic. 1999. Modeling ground water flow and radioactive transport in a fractured aquifer. *Ground Water* 37, no. 5: 770–784.

Pohll, G., and T. Mihevc. 2000. Data decision analysis: Central Nevada Test Area. Publication No. 45179. Reno, Nevada: Desert Research Institute, Division of Hydrologic Sciences.

Pohlmann, K.F., A.E. Hassan, and J.B. Chapman. 2000. Description of hydrogeologic heterogeneity and evaluation of radionuclide transport at an underground nuclear test. *Contaminant Hydrology* 44, 353–386.

Sargent, R.G. 1990. Validation of mathematical models. In *Proceedings GEOVAL '90: Validation of Geosphere Flow and Transport Models,* SKI Symposium, Stockholm, Sweden, 571–579. Stockholm: Swedish Nuclear Power Inspectorate.

Tsang, C.F. 1987. Comments on model validation. *Transport in Porous Media* 2, no. 6: 623–630.

Tsang, C.F. 1991. The modeling process and model validation. *Ground Water* 29, no. 6: 825–831.

Wen X.-H., J.E. Capilla, C.V. Deutsch, J. Gómez-Hernández, and A.S. Cullick. 1999. A program to create permeability fields that honor single-phase flow rate and pressure data. *Computer & Geosciences* 25, 217–230.

Zimmerman, D.A., G. de Marsily, C.A. Gotway, M.G. Marietta, C.L. Axness, R.L. Bras, J. Carrera, G. Dagan, P.B. Davies, D.P. Gallegos, A. Galli, J. Gómez-Hernández, P. Grindgrod, A.L. Gutjahr, P.K. Kitanidis, A.M. LaVenue, D. McLaughlin, S.P. Neuman, B.S. RamaRao, C. Ravenne, and Y. Rubin. 1998. A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by ground water flow. *Water Resources Research* 34, no. 6: 1373–1413.

Zuidema, P. 1994. Validation: Demonstration of disposal safety requires a practicable approach. In *Proceedings GEOVAL 1994: Validation Through Model Testing,* NEA/SKI Symposium, Paris, France, 35–42. Paris: Nuclear Energy Agency.